Beyond the Basics: Internet Reference Techniques

Who should attend:

Public library staff members who provide reference service on a regular basis and already have a solid background in using the Internet. **This is NOT a basic Internet class.**

Course prerequisites:

--your library must have Internet access where you can use your new skills immediately after taking the class;
--you must have basic competence in and independent experience with using a graphical web browser to navigate the Internet;
--know which helper applications/plug-ins are available on the Internet computers in your library;
--be familiar with your library's Internet policies and procedures.

Course objectives:

Knowledge objectives:

Participants will understand:

· When it is appropriate to seek information via Internet resources vs. traditional sources; the types of reference inquiries which may be readily answered using Internet resources.

· The various types of Internet finding aids and search engines.

· That various file formats for Internet documents require special software.

· Methods for maintaining awareness of new Internet sites of reference value.

Skills objectives:

Participants will learn how to:

· Identify and execute a series of steps that comprise a research strategy using the Internet.

· Conduct effective searches for information using several web-based search engines in both simple and advanced search modes.

· Critically evaluate Internet resources for authority, reliability, accuracy, objectivity and currency.

· Cite information retrieved from the Internet.

South Carolina Sites

SC Connects   www.state.sc.us/scsl/sconnect.html

SC Reference Room   www.state.sc.us/scsl/refdesk.html

SCIway   www.sciway.net

SC State Home Page   www.state.sc.us

Search Engines

AltaVista   www.altavista.digital.com

Hotbot   www.hotbot.com

Metasearch Engines

Dogpile   www.dogpile.com

Metacrawler   www.metacrawler.com

Ask Jeeves   www.aj.com

Metasites

GPO Access   www.access.gpo.gov/su_docs/dbsearch.html

HealthGate   www.healthgate.com

FindLaw   www.findlaw.com

Librarian's Index   sunsite.berkeley.edu/InternetIndex/Index.html
to the Internet

Steps for Performing Effective Web Searches

1. Know what you're looking for.

What is your question? (It helps to write it down.)
What kind of information will help you answer the question?
What organization (or type of organization) is likely to produce this information?

2. Consider whether this type of information is likely to be available over the Internet. If unlikely, consider more traditional channels of research first.

Check in library resources first. Unless you know where you are going, searching the Web is time-consuming.
Copyright-protected information? Publication date?
Book-length resources are more palatable in their "traditional" form.

3. Plan your search.

During your reference interview think of relevant keywords or phrases. (Write them down.)
Add any synonyms or alternative spellings.
Think also of broad subject categories that encompass your topic.
Think of known organizations likely to be concerned with the topic.

4. Perform your search in the most appropriate search tools.

Directory type search tool (Yahoo) and/or keyword search engine (AltaVista).
Try multiple search engines.
Use your keywords in a variety of combinations.
Read the "search hints" or help and modify your search.
Get to know the on-line sources that are needed in your community.

5. Skim the results list.

Large results list? If a likely source is not near the top, modify your search and try again.
Small results list with no relevant sites? Modify your search and try again.

6. Investigate likely sites/documents in your results list.

Repeat steps 4 and 5 as needed.

7. Before you use the information you find, evaluate it.

Who authored the information?
Is it reliable, accurate, authoritative, valid, up-to-date?

Web Search Tools: Basic Principles

Directories:

- **Examples:** *Yahoo*

- **What:** Subject-classified arrangement of sites; hierarchical arrangement--broad subject categories, subdivided into more specific topics.

- **Creation & Maintenance:** Human editors classify, arrange and (sometimes) describe and rate sites.

- **Searching:** Users can browse through an organized menu of topics and sub-topics; most directories also allow some kind of keyword search capability.

Search Engines:

- **Examples:** *AltaVista, HotBot, Excite, Infoseek.*

- **What:** Database or index of words, URLs, and other information from many sites.

- **Creation & Maintenance:** Created by software "agents" (referred to as: spiders, crawlers, or worms) which are programmed to visit and record information from sites.

- **Searching:** Users must submit a search statement or query which is run against the database according to the search logic of that search engine.

Hybrid Search Engines

- **Examples:** *Hotbot, AltaVista, Infoseek, Excite*

- **What:** To further confuse matters, some search engines also have an associated directory. These are sites that have been reviewed or rated. For the most part, these reviewed sites do not appear as the "default" when a query is made to a hybrid search engine. Instead, a user must consciously choose to see the subject section.

  Searching: User will access a search engine web site and will see the addition of a subject directory within the main search screen.

Web Search Tools: Basic Principles - Continued

Common elements of Web search tools:

- Most use an automated means to identify Web pages and other resources in order to create the database which you search. Very few use manual (or human-assisted) indexing and abstracting procedures; exceptions are several of the directory type search tools, (EX: Yahoo, Magellan, A2Z).

- Most use "relevance ranking" to weight or rank search terms. Web pages are usually retrieved and ranked according to how many of your search terms appear, how frequently they appear, and where the terms are placed within the Web site.

- Many search engines allow you to specify the number of items to be retrieved and/or a minimum relevance ranking score of items to be retrieved.

- Most search engines offer you several levels of search complexity: "simple" searching and "advanced" searching.

- All search tools provide direct links to the Web pages/resources which they index.

Variations among Web search tools:

- Size of the database or directory; how many Web sites are included.

- Types of Internet resources that are included: Web sites only, or inclusion or option to search other Internet resources (most often, USENET newsgroup postings).

- Frequency of updates: addition of new sites, deletion of information no longer available at a site, correction of changed URLs.

- Content of the database: full-text of every Web page or selected information from pages at a Web site (EX: title, header, URL only).

- Search logic employed by the search engine--how the search is executed, how terms are weighted.

Web Search Tools: Basic Principles - Continued

Search features/capabilities offered:
-natural language and/or Boolean searching
-phrase and proximity capabilities
-case-sensitive searching
-truncation or wildcard capability
-field searching

- Concept or "fuzzy" searches in addition to keyword searching (finding "related topics" as an option or supplement to your keyword search).

- The speed with which searches are conducted; how busy is the site?

- Availability and user-friendliness of search hints/tips or help.

Guides to Web Search Tools:

Search Engine Watch   www.searchenginewatch.com

An outstanding site with information and links on all aspects of search engine operations, comparisons and reviews, search tutorials, etc. Search Engine Report Mailing List: The Search Engine Report is a free, monthly newsletter about search engines and changes to Search Engine Watch. The report is sent out near the beginning of each month. Click on the link, "Free Mailing List: The Search Engine Report," to subscribe.

Internet Search Tools   lcweb.loc.gov/global/search.html

Organizes links to a large collection of Web search tools--both subject directories and search engines; also includes links to comparisons of various search engines and geographically-arranged lists of "all" Web servers.

"2nd Annual Search Engine Shoot-Out." *PC Computing* September, 1997 (v. 10 #9), pp. 196-204.   www.zdnet.com/pccomp/features/excl0997/sear/sear.html

Reports results of head-to-head tests of 4 top search engines: AltaVista, Excite, HotBot and Infoseek, with HotBot declared the winner. Also gives "search tips of the pros." Article available on-line at:

Guides to new Web sites:

Visit these sites to learn about new Web sites that have become available on the Internet.

Netscape What's New   guide.netscape.com/guide/whats_new.html

Netscape users just click on the What's New! button

Scout Report   wwwscout.cs.wisc.edu/scout/report/index.html

Includes details for subscribing to the Scout Report--an emailed announcement about excellent new web sites.

What's New on Yahoo!   www.yahoo.com/new

Good choice for finding sites relevant to current events.

Web Search Engine Capabilities

While the capabilities and features of different search engines can vary significantly, the following features are frequently available. A general explanation is given for each, but it is imperative that you consult each search engine's "Help" or "Search Tips" to learn whether the feature is offered and how to use it correctly in that search engine.

Boolean Searching: Allows you to broaden or narrow your search by using logical connectors, known as Boolean operators.
URL: www.albany.edu/library/internet/boolean.html

Examples:

prisoners **and** florida

narrows a search, requiring that *both* words appear at the sites retrieved.

prisoners **or** inmates

broadens a search, requiring that *either* or *any* of the search words appear at the sites retrieved.

prisoners **not** women

narrows a search, indicating that the first word *but not* the second word appear at the sites retrieved; note that some search engines require that the user input the words "**and not**" (AltaVista, Excite) to indicate this Boolean option.

Phrase Searching: Allows search words to be treated as a phrase (**adjacent** to each other with no intervening words). In many search engines, this is accomplished by putting quotation marks around the search words, for example: **"florida department of corrections"**

Proximity Searching: Allows searching for one word within a certain number of words of another word, thus narrowing the search. For example, in AltaVista's Advanced Search, placing the word **near** between your search words will require that those words appear within 10 words of each other, but in any order, at the sites retrieved, example: **hotbot near evaluation**

Case-sensitive Searching: Allows you to search for words in which specified letters appear in upper or lower case. For example, in AltaVista, a search for: **eXtend**   will only retrieve sites where that word appears exactly as you typed it. In most search engines, typing your search word in all lower-case is preferred, retrieving occurrences of the word regardless of case.

Web Search Engine Capabilities - Continued

Truncation Searching: Allows searching for different word endings or plurals with the use of a specified symbol. For example, in AltaVista's Advanced Search, the asterisk is used as the truncation symbol, and will retrieve the word stem. Hence a search for: **prison*** will retrieve: **prison, prisons, prisoner, prisoners**. In some search engines, words are automatically "stemmed" so that both singular and plural versions of the word will be retrieved in the search.

Field Searching: Web pages are made up of many parts or fields, such as: title, URL, text of the page, links from the page, images on the page, etc. Some search engines allow you to restrict your search to words or information found only in a specified field(s), thus narrowing your search. For example, in AltaVista, such a search would be: **title: "kids count"**

Formal Standards for Meta-Search

Daniel Dreilinger

May 6, 1996

Meta-search engines, tools that simultaneously search multiple conventional search engines and integrate the results, are becoming increasingly popular. Today at least three meta-search engines are in wide use: SavvySearch, MetaCrawler and WebCompass; many more are in development. Users report that these tools are very helpful in their Web navigation endeavors. In some cases, the search engines that are queried by meta-search engines find this behavior beneficial. Lesser known search engines enjoy the publicity and extra awareness that the meta-search tools raise. Meta-search engines also serve as additional entrances into sites whose search engines index local content only.

In other cases, as recently suggested on the robots mailing list, meta-search engines appear to work against the advertiser supported business model adopted by some of the larger search sites. Related problems that have surfaced are the increased strain on the Internet and various search engines, and reformatting of results. One solution to the advertising problem that has been suggested is propagation of advertisements produced by search engines into the meta-search results. Another solution might involve intermediate result pages which give search engines an opportunity to display advertisements for each of their links that is followed.

Ultimately it should be up to search engine providers to decide how and under what conditions their resources are used, and each will probably have a unique opinion. Perhaps these problems are best addressed with the introduction of a formal standard for meta-search tools. A standard for meta-searchers could exist as an extension to the existing robot exclusion standard, or as an entirely new mechanism (how about SavvyNotWanted.txt?) Below is a partial list of questions that I believe should be considered when designing such a standard:

- Where and when are meta-search agents welcome? (i.e., certain peak hours that should be avoided.)
- Are there maximum resource quotas that should be observed? (i.e., maximum allowable number of queries per day.)
- How much liberty may be taken in reformatting results? (i.e., bypassing or changing format of advertisements.)
- Are there other special instructions meta-search designers should follow?
- Is there a protocol for searching a fee-based service on behalf of a registered user?
- How can we avoid infinite cycles of meta-searcher queries?
- Does the standard need to be machine parsable?

This list has probably overlooked some important issues. The next step is consulting the many search engine providers and identifying their concerns.

This page is part of the <u>DISW 96</u> workshop.
Last modified: Thu Jun 20 18:20:11 EST 1996.

A Comparison of Seven Search Engines

By Eric T. Davis

Eric T. Davis is a graduate student in the <u>Kent State University School of Library and Information Science</u>, Columbus Program. This paper was written for an Online Reference Course in the Fall of 1996.

Introduction

We are a society obsessed with convenience. We go to extreme lengths to invent devices that promise a simpler or more convenient lifestyle. This paradox is exemplified in our fascination with the Internet, as well as with our attempts to index it for access purposes. The Internet is today a rapidly evolving organism that is almost completely lacking in fundamental organization. The question of whether each individual achieves a *net gain* from all the effort expended in this process lies somewhere beyond the scope of my project, but I think we all can agree on the need to somehow organize this very unstructured information resource.

Internet search tools have been created to answer this very pressing need. They are evolving rapidly - some would say more rapidly than the Internet itself. By the end of 1996, it is estimated that the Internet will consist of no less than 150 million pages, containing 50 or 60 billion words. To make matters worse, this great mass of data exists completely without any kind of bibliographical controls, standard numbering systems, or classification systems. Clearly, automated tools of some sort are necessary to sift through this mass of material (Venditto, 1996).

My own personal interest in the Internet has increased in direct proportion to the growth, power, and flexibility of the excellent search tools that have appeared over the past year or two. In my opinion, they have elevated the Web from a simply a *browsers paradise*, to a more respectable, searchable, and interesting world-wide reference source. In fact, the same critical skills that are used to locate books, journal articles, musical scores, or any other information resources, can and should be applied to finding information on the Internet (Tillman, 1996).

The Tools

Internet tools can be categorized in two very broad areas:

- Search engines (includes meta search engines & multi-treaded search engines)
- Subject catalogs/meta indexes (includes annotated directories & subject guides)

Some of the best known Internet tools can be identified as belonging to one discreet category, while others are a combination of both.

This paper will concentrate on search engines and their characteristics only. My target audience is our class, that is, sophisticated and experienced students of electronic databases, who are well aware of the established methodology used to search them. Unfortunately, a discussion of the many wonderful Internet subject guides and annotated directories will have to be done by another writer, another time.

Criteria for Choosing an Engine

Quite simply, the choice of search engine is entirely dependant on the needs and preferences of the searcher. These needs can be every bit as diverse as the Internet itself. Taking a very broad overview, search engines are the tool of choice when the searcher has a specific question in mind. They are prone to delivering very high recall, so it is imperative that they offer features that allow the searcher to narrow and limit his search. On the other hand, the subject catalogues are more appropriate for browsing the Net, and their retrieval characteristics can be described as high precision.

Search engines, if used properly, are able to match search terms with corresponding terms contained in specific Web sites. Many of the newer engines incorporate a spider or robot software to index Web sites. This automated process actually visits each new Web page and records the full text of every page (including as many as three of the page's links). Other engines may only base their indexing on the title, heading, and say the first 200 words of the body. Still others may analyze the number of links that *point to* the page being indexed, to determine its usefulness. The point is, each search engine goes about the job of indexing in a different way. The other half of the process, the front end offered to the user via the search screen form, also varies widely in terms of the operations and features engineered into the software. Some engines permit the user to key in all the necessary control language such as Boolean operators, proximity operators, and various limiting schemes. Other simply present forms with pull down menus that allow the user to select to proper limiting terms. The later technique is referred to as "form based" controls (see Comparison Table). The bottom line is that search engines rarely yield the identical results when presented with identical search terms. The user, in able to use each engine effectively, needs to understand the difference in the construction and use of each, in order to make an informed choice of product.

All search engines match the user's search terms to documents in roughly the same way (Sullivan, 1996). These are simply:

- Keywords are in the first few words of the document (keywords in title, sub-title, etc.)
- Keyword are found close to one another in a document (keyword proximity)
- Documents contain more of the query words than others (keyword frequency counts)

If this all sounds strangely familiar to DIALOG, OPAC, and electronic database searching in general, it should, the concepts aren't essentially different. However, they have been transposed and rechristened by many of the familiar search engines - much to the consternation of those of us who understand the principles involved. The best of the search engines, Alta Vista, HotBot, Infoseek Ultra, Excite, and several others do offer the searcher well-established controls that are absolutely critical for weeding the millions of sites that exist on the Web (more detail to come on this later in the next section).

Subject catalogs are actually hierarchically organized indexes of subject categories that permit the searcher to browse through lists of Web sites by subject in search of relevant information (Tyner, 1996). The analysis of sites by subject is done by humans, not computers, and therein lies both their advantage and disadvantage. First the disadvantage: the pool of indexed sites is necessarily smaller in comparison to search engines that use an automated *robot spider* to collect indexing information. However, no amount of word frequency counting or proximity calculation can compare with the interpretative ability of the human mind. So, when browsing a subject catalog, one can be assured of subject relevancy (high precision), but not comprehensiveness (high recall). What is the best answer for the poor researcher !?!

In the case of search engines, the more powerful the controls the searcher has to sort and manipulate the hits in a predictable and intuitive fashion, the better. As in all other forms of electronic querying, the user simply must take time beforehand to analyze and list as many relevant, synonymous and necessary terms as possible. The more precise the query, the more likely the material retrieved will be useful. The searcher also needs to consider the level of responses needed. To state this concept simply, the user may want to approach the subject very broadly in order to gain an idea of just how large the body of information is relevant to his topic. Or, he may want very specific, exacting information about the topic to answer questions or help to confirm a hypotheses.

The next section will address individual search engines ranked in order of preference.

Search Engines

1. AltaVista

The searcher should proceed immediately to the AltaVista Advanced Search option in consideration of the fact that this engine indexes all existing Web pages full-text (it claims 30 million). The searcher needs every control tool offered by AltaVista to avoid being hit by a tidal wave of sites. AltaVista also offers searching of News Groups on the Web. It's not unusual for an unfiltered search to yield over 100,000 hits returned for a single query - in one second! One should always head straight for the advanced search mode - or for the beta page - in any search engine. It will always provide the tools for a more controlled search.

Ever since AltaVista first exploded on the scene in December, 1995, it has been recognized as the premier search engine. It is regarded as being the most comprehensive of the search engines in terms of URLs indexed, although interestingly enough, no one seems to agree about just *how many* Web Sites are out there. At any rate, AltaVista's search results are also consistently more comprehensive than its competitors (Venditto, 1996). I concur absolutely with this conclusion.

The performance of the major search engines are similar with fairly simple searches, but as the concepts become more and more complex, the differences in engines became more apparent. The searcher can construct search phrases for AltaVista much like the phrases used in DIALOG and many other similar electronic databases. This has not always been the case for Internet search engines. Boolean, proximity searching, phrase searching, and field searching are allowed, and can be stated in the syntax that has been well established over the years (why reinvent the wheel?). Also available are the use of wildcards (an AltaVista exclusive) and case sensitivity. Examples of "good" search strings for AltaVista include (Gray, 1996):

- horses AND carriages
- "Abraham Lincoln" AND "civil war"..or.. ("Abraham Lincoln") AND ("civil war")
- ("Abraham Lincoln") AND NOT ("civil war")
- "Thomas Middleton" OR "Beaumont and Fletcher"
- (dogs OR cats) AND ("pet care")
- "William Shakespeare" NEAR Internet

Take note that the boolean NOT must be stated AND NOT, and phrases must be placed in quotes, although the parenthesis are optional. AltaVista also permits a window for the searcher to rank his search terms, a very useful device. The resulting search will be weighted to the top terms in your ranking. These user controls can help to pare down the mountain of information that AltaVista is prone to providing if not used. (Note: click on the "Comparison Table" link on left for a feature table comparison of the seven search engines discussed in this paper).

2. HotBot

HotBot is so new on the scene that few have had time to actually test and review it. It seems that once AltaVista paved the way, HotBot and several other search engines have created Internet tools that are very similar in speed and control, which also offer some unique features as well.

HotBot boasts of having indexed no fewer than 54,000,000 net sites (as of October 29, 1996), and supports the boolean AND and OR, phrase searching, limiting by date, media type, and location in its

form based menu. Once again, the experienced user should head straight for the "Expert Search" mode to gain maximum control of the 54 million options. A feature that permits the user limit by media type is unique to HotBot. With this feature, the user can access all the sites that feature specific software add-ons like Java, JavaScript, Shockwave, Acrobat, audio, or VRML viewers. This is a great way find sites to test newly downloaded software. Also, I found the Graphic layout for this page to be attractive in an austere, "generation-X" sort of way. In terms of speed, all other variables considered, all of these major engines are amazingly fast. Somehow, the program is able to search all 50 million sites in about one second.

3. *infoseek ultra*

This new engine was introduced on August 14, 1996, and offers a major improvement over its predecessor, infoseek guide, which is still very much alive. This very impressive new product also boasts of having over 50 million URLs in its index, but what really sets it apart from the others is what infoseek calls its "real-time index" of the Web (Grady, 1996). This rather obtuse phrase really means that infoseek is actually updating its index continuously. Its spider senses new and changed pages and updates the index immediately.

I must admit to a healthy doubt concerning this claim, so I put it to a test. I clicked on their *Save URL* link on the home page, and submitted all three of my personal home pages in a very short and simple process (it may have taken 25 seconds). I immediately went back to the infoseek search screen and entered appropriate search terms for my pages, and all three came up in the first ten hits! Take note all Web authors! No other search engine I tested can came close to the instantaneous refreshment that infoseek has perfected. The only other engine that is even close is AltaVista at under twenty-four hours from posting to index. This is to me a critical factor, because I feel that one of the Internet's most positive characteristics is its currency. To say that this engine is constantly the most current of all the engines is high praise indeed!

Some estimates claim that almost half of the URLs on the Web are either duplicates or dead/invalid links (INFOSEEK, 1996). Infoseek ultra has created software that filters out duplicate and/or dead links, and this too is a major feature of this engine. I have yet to get an invalid link message in any of my infoseek ultra searches. These searches are lean and accurate, with a very high "signal to noise ratio", also known as high precision.

Other useful search features include case sensitivity, proper name recognition (the search term "Junkin" *alone* sends my B.F. Junkin Home Page to near the top of the hit list), limiting search terms to particular fields, and eliminating terms with a minus sign ("-"). I would prefer more traditional syntax to execute some of these controls, but all in all, it is very difficult to find much to criticize in infoseek ultra.

4. Excite

Excite is the first engine discussed here that qualifies as both an effective Web directory organized by category and a Web search engine. It also lists 50 million indexed URLs so it can't be criticized for having a smaller pool of pages like the other Web directories. In fact, "Excite provides the fullest range of services of all the Web search sites" (Venditto, 1996). The user can search the text of at least 10,000 newsgroups, a daily news summary, opinion columns, cartoons, and Web site reviews.

Excite allows searching by **keyword or concept,** and offers searching in all the above- mentioned areas: usenet newsgroups, reviews, web documents, or classifieds. Allowable Excite search terms include (Gray, 1996):

- (illegal AND immigrant) AND NOT (Mexico)
- alien OR UFO
- alien AND NOT UFO
- football AND (rugby OR soccer)

It also offers an option to retrieve "More Like This", a kind of citation pearl-growing feature ("Query By Example" as Excite calls it), that is an essential ingredient in so many sophisticated electronic databases today. The user can pick a document that is a good match to the desired reference question, click a button next to it, and automatically reinitiate the search using the indexed search criteria of this document. This is a useful feature that seems to be unique to Excite. The fact that Excite is not only a search engine but also a Web directory, provides it with the information to make these *see also* type recommendations.

However, the tests that I performed on Excite included trying to access my three home pages using my own specified search terms. They produced some very strange results. For my first home page with the HTML Title: "Letters from the 126th Ohio Volunteer Infantry", I keyed in "126th Ohio Volunteer Infantry" and got 236 hits. My page was not in the first sixty of them. I then keyed in the complete title verbatim, and got zero hits. A little unnerved, I decided to try my second page entitled "Decedents of Johann Tobias Horine". This time, my Horine page was hit number 1 (as it should have been), but amazingly enough, my 126th page (which is a link off the Horine page) showed up as hit number 5. Go figure! Not only that, but **all** my links from the Horine page were listed in the top 7 links. One of the links, "Civil War Ohio - A special Collection" was listed even though it is a link from the first page (the 126th OVI). In addition, the Excite document summary for this link consisted of a couple of random sentences from the middle of the document. This is totally inexplicable to me, so I won't attempt it here. Suffice it to say, if I can't get predictable results when I key in my own search terms for my own pages, I tend to generally distrust the keyword matching ability of this engine across the board.

Lastly, I find the Excite screen cluttered and more than a little obtuse. Don't bother clicking on the <u>Advanced Search</u> link unless all you're after is information, because you cannot enter search terms from the advanced screen, you have to back out to the original screen to perform a search.

5. Lycos

Many veteran Web searchers have very soft spots in their hearts for Lycos, because for a while after its 1994 inception at Carnegie Mellon, it was alone in its class. After all, how can anyone dislike a search engine that was developed by a man named Dr. "Fuzzy" Mauldin? At any rate, Lycos is still quite popular, but objectively speaking, it hasn't quite kept pace with some of the newer shinier engines. It does claim an index of 68 million URLs, and their concept is to allow Internet user to:

1. "Search for specific subjects or destinations...
2. browse interesting categories...
3. [and have] a guided tour through sites of interest." (LYCOS, 1996).

Thus, Lycos strives to be all things to all people: a search engine, a subject index, and an annotated directory. I will comment only on its characteristics as a search engine.

In the tests I performed with my own URLs, Lycos performed perfectly and predictably. Generally though, Lycos is known for high recall but poor precision (Venditto, 1996). I must agree. For example, I keyed in the exact title of my 126th OVI page, and got back 364 documents, with my page right where it should be, number 1. With a search this precise, I wonder why Lycos retrieved so many other documents. In the identical search in AltaVista, I got one hit, my page. If the search terms are this precise, I think the response of the database should be equally precise. I found the Lycos response soft; if I had wanted to retrieve related documents, I would have made a more general query. This is a small point perhaps, but it makes me wonder just how many "soft" hits I would get with a more general query - probably way too many. The level of the response should match the level of the query, and this is I believe, a basic database heuristic.

The summaries of the retrieved documents are informative, with the search term bolded, a feature that would be beneficial for all engines to incorporate. Its use of boolean operators is frankly a little

Generally, Lycos retrieves lots of documents, so it's probably not the best engine for finding something quickly. It is very comprehensive, but its control language is inferior to several of the newer shinier engines listed previously.

6. Open Text

Opentext is a little secretive about the size of their index. Estimates are that it is in the range of 1.5 million URLs (Sullivan, 1996). This is considerably smaller than the 50+ million claimed by Excite, HotBot, Lycos, and Infoseek Ultra. Ironically, in the FAQ information linked from their main page, they go way out of their way to kick-around WebCrawler for only indexing 100,000 or so sites (Opentext, 1996). The truth is that of the major search engines, Open Text is next to last in index size, and WebCrawler is the only smaller one. I must say, Open Text does pick on someone its own size, *the only one!*

These concerns aside, Open Text is arguably the best-designed search site on the Web (Venditto, 1996). Open Text offers seemingly every conceivable search option. Its robot indexes each page full-text, my personal method of preference for access. It offers "power search" which can include up to five search terms and the use of five boolean operators between terms selected from pull-down menus. You can specify field searching per term: anywhere, title, summary, first heading or URL. And finally, you can specify a weighted search for up to four search terms. These options are mostly quite accommodating, but I found them to be quite linear. For example, when I entered a complex series of terms in the main menu, it only retrieved documents in which these terms occurred in the order I created.

The bottom line for Open Text is that this engine offers nice control options, but it's not nearly comprehensive enough. It is better to stick with the big indexes, and these days there are quite a few excellent ones from which to choose.

7. WebCrawler

As previously mentioned, WebCrawler has the smallest index of the major search engines, estimated at 500,000 URLs (Sullivan, 1996). It does index its sites full-text, but WebCrawler's principle criteria for selecting sites to add to the index is page popularity, or the sites that are the most well-traveled in terms of visitors. To my mind, this method would tend to yield sites that are "pop" in nature, or concerned with mainstream information. This type of construction is very well-suited to its new sponsor, America On Line. I would not look in WebCrawler for scholarly or esoteric information, however.

Another problem is that only the page titles of each retrieved URL are displayed for the searcher. This title may or may not be descriptive enough to provide intellectual access to the documents. The searcher is forced to link to each page to get a sense of its content.

If the object of your search is mainstream information, such as information on high-profile corporations, television networks, sports, or movie stars, WebCrawler should be your first choice. This is more the character of this index, and it does occupy a distinct niche. I must add however, that judicious use of control language when using the more comprehensive engines like AltaVista, HotBot, or Infoseek ultra, should enable the searcher to locate the same material.

WebCrawler is fast and easy to use. It does offer a browsable subject catalog, and in the "advanced mode" it offers boolean and proximity searching to hone your search. But, once again, WebCrawler's index is only 1% of the size of the big indexes, so I really cannot conceive of a good reason for using it as a search engine. "Compared with the newer speed merchants such as AltaVista and HotBot, WebCrawler isn't the fastest or most up-to-date search engine" (Page, 1996).

Of the subject indexes on the Web, <u>Yahoo</u> is generally regarded as the largest and best tool (Gray, 1996). If you would prefer an interesting approach to an Internet index based on the Dewey Decimal System, check out the <u>BUBL (BUlletin Board for Libraries)</u> Information Service where the URLs are divided into subject hierarchies based on Dewey.

Other very good subject indexes include:
<u>Argus Clearing House</u>
<u>Galaxy</u>
<u>Scott Yanoff's Internet Directory</u>
<u>WWW Virtual Library</u>
<u>Magellan</u> (Note: Magellan is also considered an annotated directory)
<u>Essential Links</u>

Examples of excellent meta-search engines (also known as multi-threaded search engines) include:
<u>Metacrawler</u>
<u>SavvySearch</u>
<u>Quarterdeck</u>

These search tools allow the searcher to perform a search combining the results from a variety of multiple search engines, in a customized combination specified by the searcher. The user is presented with a list of hits, and information on which search engines (i.e. AltaVista, Lycos, HotBot, etc.) they came from. The user then can simply click on any of these documents just as he would in a single engine search.

Conclusion

Before a researcher logs onto the Internet, he needs to answer a few simple question to help him determine the best type of search tool for his purposes. If he is looking for specific information, the best choice is a search engine in the order of preference as indicated in the body of this paper. If the purpose is merely to browse sites to learn what is available on the subject of interest, the subject indexes are the place to start. The meta-search engines are alluring, but theoretically at least, search engines that are comprehensive like AltaVista, HotBot, Infoseek ultra, Lycos, and Excite should yield much the same results. When using these comprehensive engines, the searcher needs to be as explicit as necessary to retrieve the level of results desired. Also, if precise information is needed, the search terms likewise need to be as precise and limiting as possible. As previously mentioned, AltaVista seems to be the best at matching the level of search terms with its level of retrieved documents, and for this and many other reasons, is my first choice for an Interenet search tool.

Bibliography

Babb, Chris. March, 27, 1996. Babb's bookmarks. Boardwatch vol. X: Issue 3.
Hypertext Document: <u>http://www.boardwatch.com/mag/96/apr/bwm8.htm</u>

Berkeley Digital Library SunSITE. September 17, 1996. Internet search tool details.
Hypertext Document: <u>http://sunsite.berkeley.edu/Help/searchdetails.html</u>

Dvorak, John C. November 7, 1996. Indexing the internet. PC Online.
Hypertext document: Http://www. Pcmag.com.dvorak.jd961007.htm.

Grady, Steve. August 14, 1996. Infoseek introduces Infoseek Ultra. Hypertext
document:Http://C%7C/Program%20Files/Netscape/Na...32.19960930112925.006d5d78@corp&number

, Library. Hypertext document: http://issfw.palomar.edu/Library/TGSEARCH.HTM

Infoseek Ultra, November 2, 1996. Comparison of world wide web search engines.
Hypertext Document: http://www.infoseek.com/doc?pg=comparison.html

Jones-Catalana, Cynthia N. November 1996. One-stop surfing. Internet World vol 7, no. 11.

LYCOS, Inc. October 23, 1996. LYCOS, Inc profile. Answer to e-mail inquiry by author.

Mitchell, Steve. May, 14, 1996. General interest resource finding tools: A review and list of those used
to build infomine. University of California, Riverside.
Hyper text document: http://lib-www.ucr.edu/pubs/navigato.html#1

Northwestern University. June 10, 1996. Evaluation of selected internet search tools.
Hypertext Document: http://www.library.nwu.edu/resources/internet/search/evaluate.html

Open Text Index. November 2, 1996. Frequently asked questions about the Open Text Index.
Hypertext document: http://index.opentext.net/main/faq.html

Page, Adam. 1996. The search is over. PC Computing Sept 12, 1996.
Hypertext document: http://www.zdnet.com/pccomp/features/960912/sub2.html

Solock, Jack. June 3, 1996. End user's corner: Site-ation pearl growing. Internic news.
Hypertext Document: http://rs.internic.net/nic-support/nicnews/archive/june96/enduser.html

_____, 1996. Searching the internert part II: Subject catalogs, annotated directories, and Subject
guides. Internic news.
Hypertext document: http://rs.internic.net/nic-support/nicnews.endusers.html.

Sullivan, Danny. The Webmaster's Guide to Search Engines and Directories. Hypertext document:
http://calafia.com/webmasters/study.html.

Tillman, Hope N. Evaluating Quality on the Net. July 10, 1996. Hypertext document:
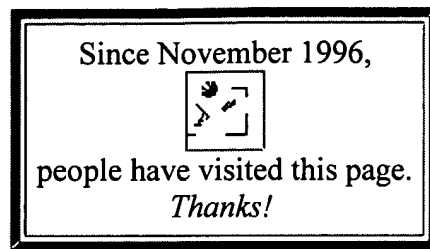http://www.tiac.net/users/hope/findqual.html

Tyner, Ross. May, 10, 1996. Sink or swim: Internet search tools & techniques. Okanagan University
College. Hypertext document: http://www.sci.ouc.bc.ca/libr/connect96/search.htm

Venditto, Gus. "Search Engine Showdown." Internet World 7.5 (May 1996): 79-86.

Webster, Kathleen and Paul, Kathryn. January 1996. Beyond surfing: Tools and techniques for searching
the web. Feliciter 42.1. pp. 48-54. Hypertext document: http://magi.com/~mmelick/it96jan.htm

Winship, Ian. "World Wide Web Searching Tools." Vine 99 (1995): 49-54.
Hypertext document: http://www.bubl.bath.ac.uk/BUBL/IWinship.html

---

Since November 1996,

people have visited this page.
Thanks!

Get Tooled Up

Metadata for the masses

Paul Miller describes Dublin Core and means by which it can be implemented.

| Main Contents | Section Menu | About Ariadne | Email Ariadne | Back Issues | Search Ariadne | Other Resources |

Table of Contents

Introduction

Metadata. The word is increasingly to be found bandied about amongst the Web *cognoscenti*, but what exactly is it, and is it something that can be of value to you and your work? This article aims to explore some of the issues involved in metadata and then, concentrating specifically upon the Dublin Core, move on to show in a non-technical fashion how metadata may be used by anyone to make their material more accessible. A collection of references at the end of the article provides pointers to some of the current work in this field.

What is metadata?

The concept of metadata predates the Web, having purportedly been coined by Jack Myers in the 1960's (Howe 1996) to describe datasets effectively. Metadata is data *about* data, and therefore provides basic information such as the author of a work, the date of creation, links to any related works, *etc*. One recognisable form of metadata is the card index catalogue in a library; the information on that card is metadata *about* a book. Perhaps without knowing it, you use metadata in your work every day, whether you are noting down the publication details of a book that you want to order, or wandering through SINES or the History Data Unit in the hope of finding a particular data set of value to your research

Metadata exists for almost every conceivable object or group of objects, whether stored in electronic form or not. A paper map from the Ordnance Survey of Great Britain, for example, has associated metadata such as its scale, the date of survey and date of publication. With products such as maps, the metadata is often clearly visible on the map itself, and is expressed using standard conventions that are easily interpretable by the experienced user (Miller 1995).

The Vale of York
Topography derived from O.S. 1:50,000 digital data. Crown copyright reserved
Coastline and Hydrology derived from Bartholomew 1:250,000 digital data
A.P. Miller '96   0 ▆▆▆▆▆▆▆▆▆▆▆▆▆▆▆ 25 km

Figure 1: a simple example of map metadata (after Miller 1996a). Click on the figure (above) to see the whole map [118Kb GIF image]

In the unfathomable maze that is the Internet, things are not always as easy. These generalised standards do not yet exist, and it can be surprisingly difficult to actually find the information for which you are searching. The current generation of search engines are undoubtedly powerful, and capable of returning a large number of suggestions in response to any search, but it is almost impossible to cut through the irrelevant suggestions to find the ones you are actually interested in. A search for *Ariadne* on Alta Vista, for example, found 5,468 references, and returned 3,000 links. On the first page of links, there was a pointer to Issue 3, but nothing else relevant to *my* needs turned up until the very bottom of the third page. In this case, it was fairly straightforward to distinguish between the (relevant);

Ariadne: Issue 2 - Contents

- Contents Page for Issue 2. Welcome to issue 2 of Ariadne on the Web, the World Wide Web version of the magazine for the discerning UK Library and...
- http://www.ukoln.ac.uk/ariadne/issue2/contents.html - size 6K - 25 May 96

and the (irrelevant?);

Ariadne

- Ariadne --- A further development. 9th semester in Computer Science. by: Henning Andersen. Jan M. Due. Peter D. Fabricius. Flemming Sørensen. Supervisor:..
- http://www.iesd.auc.dk/general/DS/Reports/1989/ariadneFurther.abstract.html - size 1K - 28 Jun 94

This simple example illustrates some of the problems with finding information on the Web. It is perhaps analogous (or perhaps not!) to a paper-based list of contacts which, rather than being sorted conventionally by surname, is sorted simultaneously by the contents of *every* field (surname, company, street, *etc*). Of course, when you attempt to look up an address in this contact list, you have no way of knowing which field the result is coming from. Assuming you wish to contact our esteemed web editor to offer an article for *Ariadne* (hint!) and search for his surname (Kirriemuir), you don't *really* know whether the result you have found is really him, or part of the address of some long-forgotten relative from a small Scottish town just west of Forfar.

To make your contact list useful, you need some metadata to describe what each string of text relates to (*ie* Kirriemuir is a SURNAME or Kirriemuir is a TOWN).

Most applications are, of course, more complex than this, but it is at least possible to demonstrate the principles using this simple case study. How, then, are the 'experts' currently approaching the description of metadata?

Existing approaches to metadata

A large number of standards have evolved for describing electronic resources, but the majority are concerned with describing very specific resources, and often rely upon complicated subject-specific schema that make either widespread adoption or easy accessibility to these records unlikely. Rachel Heery (forthcoming) offers a review of some of the major metadata formats in a forthcoming article.

In an environment such as the traditional library, where cataloguing and acquisition are the sole preserve of trained professionals, complex metadata schemes such as MARC (MAchine Readable Catalogue) are, perhaps, acceptable means of resource description. In the more chaotic online world, however, new resources appear all the time, often created and maintained by interested individuals rather than large centrally funded organisations. As such, it is difficult for anyone to easily locate information and data of value to them and the large search engines - with all their faults - are often the only means by which new information may be found.

In such an environment, there is an obvious requirement for metadata, but this metadata must be of a form suitable for interpretation both by the search engines and by human beings, and it must also be simple to create so that *any* web page author may easily describe the contents of their page and make it immediately both more accessible and more useful. As such, compromises must be made in order to provide as much useful information as possible to the searcher while leaving the technique simple enough to be used by the maximum number of people with the minimum degree of inconvenience.

The expert approach

A large number of techniques exist for the description of resources in an electronic medium, ranging from the various flavours of MARC (British Library 1980, Library of Congress 1994, Heery forthcoming) used in library cataloguing to the more specialised Directory Interchange Format (DIF) which provides metadata for satellite imagery and the like (GCMD 1996).

Developments such as the Text Encoding Initiative (TEI) have gone a long way towards allowing a standardised description of electronic texts, and the ongoing review of the US National Spatial Data Infrastructure (NSDI) will hopefully succeed in realising a similar scheme for the complex issues involved in describing spatial data. In the United Kingdom, the provisionally named National Geospatial Database (Nanson *et al* 1995) is aiming to increase the integration between governmental and non-governmental spatial data holdings, and careful thought will need to be given to the construction of rational metadata schemes for this project over the next year or two.

Each of these formats has been developed to operate within a narrowly defined field of work, and is poorly suited to the description of a wider range of resources. Many of these existing metadata schemes are also extremely complex, and are geared towards creation by experts and interpretation by computers, rather than both creation and interpretation by as wide a range of interested parties as possible.

In cutting through the morass of existing - and often conflicting - metadata approaches, the work of eLib projects such as ROADS, ADAM *et al* will be well worth watching, as will the efforts of the Arts & Humanities Data Service (AHDS) to create a pan-subject metadata index that encompasses the current AHDS projects for Archaeology, History, Text and the Performing Art s, as well as any future projects. It is interesting to note that several of these projects (ADS, ADAM) have already adopted a form of Dublin Core description for at least some of their pages. As with this document, Dublin Core metadata is often stored in the <HEAD> </HEAD> area of a Web page, and may be viewed simply by selecting View... | Document Source from your Web browser's menu bar.

The search engine approach

Recognising the need for a means by which searches may be better tailored to actual user interests, a number of the current search engines have begun to include the ability to make use of the HTML <META> tag in Web documents. *Alta Vista*, for example, makes use of DESCRIPTION and KEYWORDS

qualifiers to the <u>META</u> tag in order to index a given page. The DESCRIPTION is returned in response to a search, rather than the default (but usually far less useful) first couple of lines of text.

eg

```
<META NAME="description" CONTENT="The most useful paper on metadata ever written">
<META NAME="keywords" CONTENT="Dublin Core, metadata">
```

in the <HEAD> area of this document would cause *Alta Vista* to return the following in response to a search on any of the words stored in either DESCRIPTION or KEYWORDS;

<u>Metadata for the masses</u>

- The most useful paper on metadata ever written.
- <u>http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses/</u> - size 51K - 9 Sept 96

The Dublin Core

Notably different from many of the other metadata schemes due to its ease of use and interpretability is the so-called Dublin Core Metadata Element Set, or <u>Dublin Core</u>. This approach to the description of 'Document Like Objects' is still under development, and is the focus of a great deal of activity worldwide as researchers work to produce the most useful model they can, capable of describing the majority of resources available on the Internet as a whole, and suitable for inserting into a wide range of file types from the simple HyperText Markup Language (<u>HTML</u>) of the Web to Postscript files and other image formats (*eg* <u>Knight 1996</u>, <u>Beckett 1996</u>). Despite the emphasis of this, and other, papers (<u>A.P. Miller 1996*b*</u>, <u>E. Miller 1996*a*</u>, <u>E. Miller 1996*b*</u>, <u>Weibel 1996</u>) on the HTML implementation of Dublin Core, readers should remember that the concepts are equally applicable to virtually any other file format. In the case of this article, the HTML implementation is stressed because it is felt that this is the area in which the underlying concepts may most easily be demonstrated, and because it is in the provision of metadata for the many thousands of personal pages out on the Web that a structure such as Dublin Core may most rapidly make an impact of value to readers of *Ariadne*. With luck, once you have followed the examples here and filled your text web pages with Dublin Core metadata, you will then feel both sufficiently enthused and competent to further explore the <u>references</u> in order to add metadata to your more complex file formats.

As Dempsey argues (<u>1996*b*</u>), Dublin Core metadata descriptions exist between the crude metadata currently employed by search engines and the complex mass of information encoded within records such as those for MARC or the Federal Geographic Data Committee (<u>FGDC 1994</u>).

The Core Element Set

The Dublin Core itself consists of thirteen core elements, each of which may be further extended by the use of SCHEME and TYPE qualifiers;

Element Name	Element Description
Subject	The topic addressed by the object being described
Title	The name of the object
Author	The person(s) primarily responsible for the intellectual content of the object
Publisher	The agent or agency responsible for making the object available
OtherAgent	The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
Date	The date of publication
ObjectType	The genre of the object, such as novel, poem, or dictionary
Form	The data format of the object, such as Postscript, HTML, *etc*
Identifier	String or number used to uniquely identify the object
Relation	Relationship between this and other objects
Source	Objects, either print or electronic, from which this object is derived
Language	Language of the intellectual content
Coverage	The spatial locations and temporal duration characteristic of the object

Table 1: The fields of the Dublin Core Metadata Element Set

In creating metadata for insertion into Web pages, the HTML <META> tag is used to place the description within the page's < HEAD> </HEAD> area, as shown below;

```
<!DOCTYPE HTML PUBLIC "-IETF//DTD HTML 2.0//EN">

<HTML>
<HEAD>

<TITLE>Metadata for the masses</TITLE>

<META NAME="package" CONTENT="(TYPE=begin) Dublin Core">

<META NAME="DC.title" CONTENT="(TYPE=long) Metadata for the masses: what is it, how
can it help me, and how can I use it?">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#title">

<META NAME="DC.title" CONTENT="(TYPE=short) Metadata for the masses">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#title">

<META NAME="DC.subject" CONTENT="(SCHEME=keyword) Dublin Core, Metadata, Warwick
Framework, Resource Description, Resource Discovery">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject">

<META NAME="DC.author" CONTENT="(TYPE=name) Paul Miller">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=email) A.P.Miller@newcastle.ac.uk">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=postal) University Computing Service
University of Newcastle  Newcastle upon Tyne  NE1 7RU  UK">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=phone) +44 191 222 8212">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">
<META NAME="DC.author" CONTENT="(TYPE=fax) +44 191 222 8765">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=affiliation) University of Newcastle upon
Tyne">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">
```

```
<META NAME="DC.author" CONTENT="(TYPE=affiliation) Archaeology
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=homepage) http://www.ncl.ac.uk/~napm1/">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.publisher" CONTENT="(TYPE=name) Ariadne">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#publisher">

<META NAME="DC.publisher" CONTENT="(TYPE=email) ariadne@ukoln.bath.ac.uk">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#publisher">

<META NAME="DC.publisher" CONTENT="(TYPE=homepage)
http://www.ukoln.ac.uk/ariadne/">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#publisher">

<META NAME="DC.date" CONTENT="(TYPE=creation) (SCHEME=ISO31) 1996-09-02">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date">
<LINK REL=SCHEMA.iso31 REFERENCE="ISO 31-1:1992 Quantities & Units -- Part 1: space
& time">

<META NAME="DC.date" CONTENT="(TYPE=current) (SCHEME=ISO31) 1996-09-09">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date">
<LINK REL=SCHEMA.iso31 REFERENCE="ISO 31-1:1992 Quantities & Units -- Part 1: space
& time">

<META NAME="DC.form" CONTENT="(SCHEME=imt) text/html">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#form">
<LINK REL=SCHEMA.imt HREF="http://sunsite.auc.dk/RFC/rfc/rfc1521.html">

<META NAME="DC.identifier" CONTENT="(TYPE=url)
http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses/">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#identifier">

<META NAME="DC.relation" CONTENT="(TYPE=IsChildOf) (IDENTIFIER=url)
http://www.ukoln.ac.uk/ariadne/issue5/">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#relation">

<META NAME="DC.language" CONTENT="(SCHEME=iso639) en">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#language">
<LINK REL=SCHEMA.iso639 REFERENCE="ISO 639:1988 Code for the representation of names
of languages">

<META NAME="package" CONTENT="(TYPE=end) Dublin Core">

</HEAD>


<BODY>
...{body of document}...
```

In writing metadata such as this, the user may include as many of the elements from Table 1 as necessary, and each of these fields may be repeated several times in order to describe all relevant details. In the example above, elements such as Coverage and ObjectType have not been used at all, while those such as Author and Publisher have been used several times.

As Beckett (1996) notes, the use of case (ABC... as opposed to abc...) and whitespace (A B C... as opposed to ABC...) is not strictly defined within the Dublin Core, and may be modified to suit individual user and project requirements.

While not formally part of the Dublin Core definition, a recognised 'good practice' is evolving, whereby the Dublin Core element name is given in lower case, preceded by an identifier in upper case to denote that the element is from Dublin Core (DC.author, rather than DC.AUTHOR, dc.AUTHOR, DC.Author, etc). Also, META, NAME, CONTENT, TYPE and SCHEME should be given in upper case, while the values of each should normally be given in lower case (or a mixture of the two, where proper names etc are involved).

At the most basic, a Dublin Core entry coded within HTML should therefore take the form;

```
<META NAME="DC.element name" CONTENT="value of element">
```

eg

```
<META NAME="DC.author" CONTENT="Paul Miller">
```

Note the initial '<' and the final '>', as well as the use of " " to enclose the values of NAME and CONTENT.

Use of the <LINK> tag

Although undoubtedly easier for the casual viewer to understand than many metadata schemes, the Dublin Core still presents scope for ambiguity in understanding, both of the core elements themselves and in the many SCHEMEs involved in adding extra information.

The solution adopted for overcoming these ambiguities is to include a reference to further information through the HTML <LINK> tag (Weibel 1996, A.P.Miller 1996*b*). For each occurence of a Dublin Core element, a <LINK> is provided to the definition of that element on the Dublin Core page at http://purl.org/metadata/dublin_core_elements, and for each use of a SCHEME a link is provided to an on- or off-line definition of the syntax used within that scheme.

e.g.

```
<META NAME="DC.identifier" CONTENT="(TYPE=url)
http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses/">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#identifier">
```

shows a simple use of the Dublin Core element, Identifier, with a <LINK> to its definition, while

```
<META NAME="DC.language" CONTENT="(SCHEME=iso639) en">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#language">
<LINK REL=SCHEMA.iso639 REFERENCE="ISO 639:1988 Code for the representation of names
of languages">
```

illustrates a use of the Dublin Core element, Language. As this example includes the use of a SCHEME, an extra <LINK> is included to a definition of this schema.

A <LINK> pointer to further information may take the form of a REFERENCE to an offline source or an HREF to another web page.

eg

```
<LINK REL=SCHEMA.iso639 REFERENCE="ISO 639:1988 Code for the representation of names
of languages">
```

```
<LINK REL=SCHEMA.imt HREF="http://sunsite.auc.dk/RFC/rfc/rfc1521.html">
```

SCHEMEs and TYPES

In order to better describe the resource, the basic thirteen elements may be further enhanced by the use of SCHEME and TYPE qualifiers. As special cases, OtherAgent also has a Role qualifier, and Relation an Identifier.

The SCHEME qualifier identifies any widely recognised coding system used in the description of a specific Dublin Core element, and allows a degree of consistency and standardisation to be introduced to Dublin Core records. Instead of describing (in the Form element) a web page as being "a web page", "HTML" or "HyperText Markup Language", for example, it is far easier and more consistent to use the existing

Internet Media Types (IMT) and describe it as "text/html". In Dublin Core's HTML syntax, this would be represented as;

```
<META NAME="DC.form" CONTENT="(SCHEME=imt) text/html">
```

and should also be provided with the necessary <LINK>s, as discussed above.

```
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#form">
<LINK REL=SCHEMA.imt HREF="http://sunsite.auc.dk/RFC/rfc/rfc1521.html">
```

A SCHEME should *only* refer to the name of an existing coding system such as the Internet Media Type (IMT), or the International Standards Organisation standard on dates (ISO31), and should *not* be used for identifying, for example, that a use of the Author element is referring to a name, e-mail address, or whatever. For tasks such as this, the TYPE qualifier should be used. This suggestion differs from that given in the most comprehensive list of SCHEMES and TYPES currently available (Knight & Hamilton 1996), but appears to create a more logical use of the two qualifiers.

Knight & Hamilton (1996) suggest including the vast majority of qualifiers to a metadata entry within SCHEME and only use TYPE in a few cases. This author would suggest a different division, whereby only references to coding schemes appear in SCHEME and most other qualifiers appear in TYPE. As a simple rule of thumb, **if a <LINK> can be included to an on- or off-line definition, then it is a SCHEME and if not, it is a TYPE.** An early implementation of this model was produced by the author (1996*b*), and the beginnings of a second may be seen evolving at http://www.ncl.ac.uk/~napm1/ads/DC_scheme_type.html, where a comprehensive list of SCHEMES and TYPES will soon be available, along with guidance on usage for each.

The TYPE qualifier, then, is mainly used where a Dublin Core element occurs more than once in a metadata description. You may, for example, use the Author element several times in order to provide name, address and telephone information. In a case such as this, the TYPE qualifier would be used to differentiate between each occurrence of Author.

eg

```
<META NAME="DC.author" CONTENT="(TYPE=name) Paul Miller">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=email) A.P.Miller@newcastle.ac.uk">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=postal) University Computing Service
University of Newcastle Newcastle upon Tyne NE1 7RU UK">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=phone) +44 191 222 8212">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=fax) +44 191 222 8765">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=affiliation) University of Newcastle upon
Tyne">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=affiliation) Archaeology Data Service">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">

<META NAME="DC.author" CONTENT="(TYPE=homepage) http://www.ncl.ac.uk/~napm1/">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#author">
```

Note that TYPES and SCHEMES may be used several times within a Dublin Core description in the same

manner as the Core Elements themselves. In the example above, `Affiliation` appears twice in order to
effectively describe the affiliations affecting work on this project.

Extending the Dublin Core

Even with the great flexibility afforded by SCHEMES and TYPES, the <u>thirteen elements</u> of the Dublin Core are not capable of describing all eventualities. If the core element set were extended in order to attempt this, it would rapidly become large and unwieldy, and ultimately one of the incomprehensibly complex metadata schemes that Dublin Core was created to avoid.

The currently held view of Dublin Core is that it should *not* be directly extended itself, but that any necessary extensions should be included in a separate 'package', as proposed in the Warwick Framework (<u>Lagoze *et al* 1996</u>). Descriptions stored within this new 'package' may then either be from a totally different metadata scheme, such as <u>DIF</u> or <u>FGDC</u>, or they may be simple extensions to the thirteen Dublin Core elements, and described in a Dublin Core-like syntax.

In the same way as the package of metadata known as the Dublin Core is enclosed within

```
<META NAME="package" CONTENT="(TYPE=begin) Dublin Core">

...

<META NAME="package" CONTENT="(TYPE=end) Dublin Core">
```

so should any other package of metadata be denoted. Where the metadata scheme used is Dublin Core-like in syntax, a form for element names similar to the `SCHEME.element name` (*eg* `DC.author`) of Dublin Core should also be used.

eg

```
<META NAME="package" CONTENT="(TYPE=begin) Dublin Core">

...Dublin Core metadata in here...

<META NAME="package" CONTENT="(TYPE=end) Dublin Core">

<META NAME="package" CONTENT="(TYPE=begin) ahdsDescriptor">

<META NAME="AD.precision" CONTENT="(TYPE=spatial) (TYPE2=recorded) 2">
<LINK REL=SCHEMA.ad
HREF="http://www.ncl.ac.uk/~napml/ads/ahds_descriptor_elements#precision">

<META NAME="package" CONTENT="(TYPE=end) ahdsDescriptor">
```

What the future holds...

Given rapid changes both in metadata and in the Web itself, it is difficult to predict exactly what the future holds, but for the Web/HTML version of Dublin Core described here to be most useful, the following developments need to be pursued:

HTML

The current practice of inserting Dublin Core metadata within HTML's <<u>META</u>> tag certainly works, but enhancements to the existing definition of this tag should be encouraged in order to enable more legible representations whereby the current

```
<META NAME="DC.author" CONTENT="(TYPE=email) A.P.Miller@newcastle.ac.uk">
```

might be replaced by

```
<META NAME = "DC.author"
      TYPE = "email"
      CONTENT = "A.P.Miller@newcastle.ac.uk">
```

Whilst the latter form is accepted by the current generation of Web browser, it breaks the Document Type Description (DTD) for HTML, and therefore does not pass the majority of HTML validation tools currently used by Web authors.

Metadata creation

At present, although tools exist for the creation of metadata conforming to some of the more complex schemes, Dublin Core-style metadata must be entered by hand. Work is currently underway within projects such as the European-funded <u>DESIRE</u> (McDonald *pers comm*) to investigate means by which much of this metadata creation may be automated (<u>McDonald 1996</u>). Such automation will undoubtedly make the creation and upkeep of useful metadata more straightforward, and therefore hopefully more commonplace.

Search Engines

As discussed <u>above</u>, many of the web search engines allow the inclusion of limited metadata within the <HEAD> </HEAD> area, but this metadata is only fully used if it is in the syntax recommended for that particular engine. While representatives of several of the search engine producing companies are involved in Dublin Core development, none has yet modified their software to make full use of Dublin Core-compliant web pages. Such a development cannot be far off in happening.

Conclusion

The world of digital metadata is a complex one, currently in a state of rapid flux. As I sit in sunny Newcastle typing the last of this paper, e-mail messages continue to arrive from various lists that threaten to force a rethink of my ideas. With deadlines looming, and demonstrating a remarkable degree of willpower, I ignore these latest ideas in order to actually get this article finished in time.

As such, it is impossible to say that the implementation of Dublin Core demonstrated here is exactly the one that will be recommended six months down the road, but given all the hard work that has gone into deriving the current offering any evolution is likely to be slight. The next stage is to continue exploring different uses of the Dublin Core idea, and to approach standards bodies with a view to ratifying something in the near future.

As exactly the type of person for whom Dublin Core could offer so much, it would be extremely useful if Ariadne readers could begin to implement Dublin Core metadata in their web pages, and report back on any of the shortcomings that they discover. If you start now, you'll be a part of a growing and exciting trend, whereby all the *data* available out on the Web might actually become *information*, and therefore of use to the wider community.

A selection of useful references

Not all of these references are actually cited in the article, but they do form a useful introduction to some

Beckett, D.J., 1996, *Using Dublin Core Metadata*, Draft **0.1**, URL: http://www.hensa.ac.uk/pub/metadata/dc-encoding.html.

Beckett, D.J., Knight, J., Miller, E. & Miller, A.P., forthcoming, *A guide to implementation of the Dublin Core Element Set*, URL: http://www.ncl.ac.uk/~napm1/ads/D C_implementation.html.

British Library, 1980, *UKMARC manual* **2nd edition**, British Library: London.

Burnard, L., Miller, E., Quin, L. & Sperberg-McQueen, C.M., 1996, *A syntax for Dublin Core metadata: Recommendations from the second metadata workshop*, URL: http://info.ox.ac.uk/~lou/wip/metadata.syntax. html.

Day, M., 1996, *The UKOLN Metadata page*, URL: http://www.ukoln.ac.uk/metadata/.

Dempsey, L., 1996*a*, Meta Detectors, *Ariadne* **3**, URL: http://www.ukoln.ac.uk/ariadne/issue3/metadata/.

Dempsey, L., 1996*b*, ROADS to Desire: Some UK and Other European Metadata and Resource Discovery Projects, *D-Lib Magazine* **July/August**, URL: http://www.ukoln.ac.uk/dlib/dlib/july96 /07dempsey.html.

Dempsey, L. & Weibel, S.L., 1996, The Warwick Metadata workshop: a framework for the deployment of resource description, *D- Lib Magazine* **July/August**, URL: http://www.ukoln.ac.uk/dlib/dlib/july96/0 7weibel.html.

FGDC, 1994, *Content standards for digital geospatial metadata*, Federal Geographic Data Committee, **8 June**.

Global Change Master Directory, 1996, *Directory Interchange Format (DIF) Writer's Guide*, version **5**, URL: http://gcmd.gsfc.nasa.gov/difguide/difman.html< /A>.

Greenstein, D., 1996, AHDS: Arts & Humanities Data Service, *Ariadne* **4**, URL: http://www.ukoln.ac.uk/ariadne/issue4/ahds/.

Heery, R., forthcoming, Review of metadata formats, *Program* **30/4**.

Howe, D., 1996, *Free on-line Dictionary of Computing (FOLDOC)*, URL: http://wombat.doc.ic.ac.uk/.

IFLA, 1996, *Metadata Resources*, URL: http://www.nlc-bnc.ca/ifla/II/metadata.htm.

Knight, J., 1996, *MIME implementation for the Warwick Framework*, URL: http://www.roads.lut.ac.uk/MIME- WF.html.

Knight, J. & Hamilton, M., 1996, *Dublin Core sub-elements*, URL: http://www.roads.lut.ac.uk/Metadata/DC-SubElements.html.

Lagoze, C., Lynch, C.A. & Daniel Jnr., R., 1996, *The Warwick Framework: a container architecture for aggregating sets of metadata*, URL: http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR96- 1593?abstract=warwick.

Library of Congress, 1994, *USMARC format for bibliographic data including guidelines for content negotiation*, Network Development and MARC standards office, Library of Congress: Washington DC.

Lock, G. & Stancic, Z., (Eds.), 1995, *Archaeology and Geographical Information Systems: A European Perspective*, Taylor & Francis: London.

McDonald, T., 1996, *Welcome to the metatest site*, URL: http://www.netskills.ac.uk/staff/m cdonald/metadata/index.html.

Medyckyj-Scott, D., Newman, I., Ruggles, C. & Walker, D., 1991, *Metadata in the Geosciences*, Group D Publications, Ltd: Loughborough.

Miller, A.P., 1995, How to look good and influence people: thoughts on the design and interpretation of an archaeological GIS, in Lock & Stancic, (Eds.), pp. 319-333.

Miller, A.P., 1996*a*, The York Archaeological Assessment: an investigation of techniques for urban deposit modelling utilising Geographic Information Systems, unpublished doctoral thesis, Department of Archaeology, University of York.

Miller, A.P., 1996*b*, *An application of Dublin Core from the Archaeology Data Service*, URL: http://www.ncl.ac.uk/~napm1/ads/metadata.html .

Miller, E.J., 1996*a*, *An approach for packaging Dublin Core metadata in HTML 2.0*, URL: http://www.oclc.org:504 6/~emiller/publications/metadata/minimal.html.

Miller, E.J., 1996*b*, *Issues of document description in HTML*, URL: http://www.oclc.org:5046/~ emiller/publications/metadata/issues.html.

Nanson, B., Smith, N. & Davey, A., 1995, What is the British National Geospatial Database?, *AGI'95 Conference Proceedings*, pp. 1.4.1-1.4.5. Reproduced on the WWW at URL: http://www.ordsvy.gov.uk/osinfo/ge neral/agi95/nansmit.html.

Richards, J.D., 1996, *The Archaeology Data Service*, URL: http://intarch.ac.uk/ahds/welcome.html.

Schwartz, M., 1996, *Report of the W3C Distributed Indexing/Searching workshop*, URL: http://www.w3.org/pub/WWW/Search/9605-Indexing-Workshop/.

Weibel, S., 1996, *A proposed convention for embedding metadata in HTML*, URL: http://www.oclc.org:5046/~weibel/html-meta.html.

Weibel, S., Godby, J., Miller, E. & Daniel, R., 1995, *OCLC/NCSA Metadata Workshop Report*, URL: ht tp://www.oclc.org:5047/oclc/research/publications/weibel/metadata/dublin_core_report.html.

Weibel, S. & Miller, E., 1996, Dublin Core Metadata Element Set WWW homepage, URL: http://purl.org/metadata/dublin_core.

Acknowledgements

Enhanced with 100%

Dublin
C re
Metadata

(select View... | Document Source to see the metadata)

Main Contents | Section Menu | About Ariadne | Email Ariadne | Back Issues | Search Ariadne | Other Resources

Material on this page is copyright Ariadne/original authors. This page last updated on September 11th 1996

AltaVista is a very large database (30 million+ sites) with powerful features in its Advanced Search for refining searches and extracting the documents you want. It offers two levels of searching: **Simple** and **Advanced.** The Simple Search does not offer Boolean logic.

AltaVista basic tips:

Simple searching
Uses *phrase searching, truncation, field* and a basic joining together of words using the symbols + and -. To find an article on pet care, you might try the query **dog cat pet +care.** AltaVista will look for all pet care articles relating to dog or cat. To find a recipe for oatmeal raisin cookies without nuts try **oatmeal raisin cookie-nut* -walnut***

Advanced searching
Advanced search is for very specific queries and not for beginning searcher. Almost everything you need to do can be done more quickly and with better results through the simple form, where AltaVista controls the ranking. Use the Advance Searching feature if you need to find documents within a certain range of dates or if you have to do some complex Boolean searches. **Remember, when you use the advanced search form, you control the ranking and if the ranking field is left blank, no ranking will be applied and the results will be in no particular order.** The + and - operators do not work when using the Advanced search form.

Phrase searching
Enclose terms to be searched as a phrase in quotes. "American Dietetic Association"

Truncation
Right-hand truncation with *   femin*   retrieves   feminine, feminist, feminism, etc.

Uses Boolean logic
Uses **AND, OR, AND NOT** as well as **NEAR.**

Dates
You can restrict an Advanced Search to find only documents last modified during a specific time frame. When entering To and From dates, use the format **dd/mmm/yy,** where dd is the day of the month, mmm is the name of the month, and yy is the last two digits of the year. Be sure to use the name of the month instead of a number; this eliminates ambiguity between date formats in different countries. For example, use **09/jan/96.**

Field searching

> **url:** Requires characters to be in the URL or address of a site. Use **url:altavista** to find all pages on all servers that have the word AltaVista in the host name, path, or filename-- the complete URL, in other words.
>
> **image:** Finds images containing the words to specify of the "image" field describing the image. Use **image:elvis** to find pages with images called Elvis. Since this is usually an 8 character filename assigned by the programmer, remember to only use one word with the image field.
>
> **domain:**domainname Finds pages within the specified domain. Use **domain:de** to find pages from Germany, or use domain:org to find pages from organizations.
>
> **Internet Country Domains (By Name)**
> **URL: www.edepot.com/irname.html**
>
> Examples:

.org	non-profits	.net	network
.us	united states	.mil	military
.com	commercial	.edu	educational institution
.gov	federal government		

.AT	Austria	.IE	Ireland
.AU	Australia	.IT	Italy
.BR	Brazil	.JP	Japan
.CA	Canada	.NL	The Netherlands
.CH	Switzerland	.PT	Portugal
.DE	Germany	.SE	Sweden
.ES	Spain	.TW	Taiwan
.FR	France	.UK	United Kingdom
.HK	Hong Kong	.ZA	South Africa

> **title:**text Finds pages that contain the specified word or phrase in the page title (which appears in the title bar of most browsers). The search **title:elvis** would find pages with Elvis in the title.

Case sensitivity
Capitals retrieve only matching capitals in documents. Lower case retrieves upper or lower and is always safe.

Hotbot is another very large database (54 million pages) with considerable potential to refine searches. It lacks truncation and limiting to title field, but has optional Boolean logic and phrase searching. Because Hotbot also offers a forms-formatted option, and the option of using +requires/-excludes, it appears somewhat harder to use than it really is.

The person's name search offers flexibility unequaled in any other general Web searching tool.

Hotbot permits geographical, media-type, and domain searching not available in the other good search tools, and has many technical search possibilities invaluable to a Web expert.

Hotbot basic tips:

Simple searches
Uses *phrase searching*, and *pull-down menu* narrow your search.

Pull-down menu options include:

> **All the words**: The result of this type of search will contain at least one instance of each word in every page returned, but not necessarily in the order that you typed them.

> **Any of the words**: Selecting any of the words tells HotBot to find pages that contain one or more of the words or phrases that you typed.

> **The exact phrase**:
> > 1. Use quotation marks: "American Dietetic Association" **OR**
> > 2. If you select "the exact phrase," be sure to omit quotation marks.

> **The person**: With "the person" specified, Zeppo Marx retrieves Zeppo Marx, Marx, Zeppo, Mr. Zeppo Marx

> **Links to this URL**: Shows how many people are linked to any web site.

> **The Boolean expression**: This selection allows you to enter "advanced" searches directly as text, instead of using our Modify panel.

Case sensitivity
Only "Interesting case" capitals are respected (i.e., imbedded within words). Initial capitals and words all in capitals are treated as lower case. "**W rld War II**" retrieves both **world war II** and **W rld War II neXt** will not retrieve next but will match what you asked for.

Hotbot - Continued

NO truncation
NO right-hand or other truncation.
Do NOT try to use *
To search variant spellings, synonyms, or equivalent aspects of a word, use Boolean logic and separate the terms by OR.

Boolean operators
Uses Boolean operators and nesting.

Field searching

> **domain:** Restricts a search to the domain selected. Domains can be specified up to three levels deep (**.com, intel.com,** or **support.intel**.com).

> **feature:** Limits your query to pages containing the specified feature. Most of these controls are also available under the Media Type panel. The name can be any of the following:
> feature:acrobat   Detects Acrobat files
> feature:applet   Detects embedded Java applets
> feature:audio   Detects a range of audio formats

> **linkdomain:** Restricts a search to pages containing links to the specified domain. For example, linkdomain:hotbot.com finds pages that point to HotBot.

> **title:** This searches for pages containing the given word in their titles between the HTML tags. Any additional words with this marker could be found anywhere within the text of a document, including, but not limited to, the title.

> One should note the spacing after a colon when using a meta tag. For example, "title:[word]" is equivalent to one word, and "title: [word]" is equivalent to two words.

Date meta words

> **after:**[day]/[month]/[year] Restricts a search to documents created or modified after the specified date (e.g., currents AND after:30/6/96).

> **before:**[day]/[month]/[year] Restricts a search to documents created or modified before the specified date (e.g., "cyber crime" AND before:30/6/96).

> **within:**number/unit Restricts a search to documents created or modified within the last specified time period (e.g., (pet +care) AND within:3/months). Units can be days, months, or years.

Dogpile Searches:

The Web: Yahoo!, Lycos' A2Z, Excite Guide, GoTo.com, PlanetSearch, Thunderstone, What U Seek, Magellan, Lycos, WebCrawler, InfoSeek, Excite & AltaVista.
Usenet: Reference, Dejanews, AltaVista and Dejanews' old Database.
FTP: Filez and FAST FTP Search. (Only the first word will be passed on to FTP Search.)
Weather: Enter in any City, State or Zipcode in the world.
Stock Quotes: Enter Tickers or Company Name.
Business News: Search for Business News. Africa News , Agence France , M2 Airlines, Asiainfo, Business Wire , Canadian Corp , Content Factory, Fednet, Infolatina, Inter Press, Interactive Sports, Itar-Tass, M2, Phillips, PR News, PIO, Resource News, SABI, UPI, UPI, US Newswire, Washington Tech, WENN, Xinhua.
Other News Wires: Yahoo News Headlines, Excite News and Infoseek NewsWires.

Search Syntax:

You may use the proximity and Boolean operators AND, OR, NEAR, and NOT to combine words and phrases. NEAR will be substituted with AND for those engines which do not support its use. If you use NEAR the engines which support its use will be searched first. NOT and the following word will be deleted if the engine does not support its use. OR is not fully supported since not all the search engines included support a mixed use of AND and OR. This is not a limitation for MetaFind however.

Using no connector, AND will be assumed. Thus the search: Free and Mac and Software and Free Mac Software are the same.

You may also use quotes and parentheses. However note that not all search engines support their use. For those which do not support their use, they will automatically be removed.

The FTP search engines only take one word as the query (i.e. a file name or part of a file name) Make sure that the first word in your query is the file name you want to find. You may still add other search terms if you also are searching the web or USENET also.

What will Happen when I Press "Fetch"?

Arfie searches three search engines at a time. The requests are put out in parallel and are displayed as they come back.

If Arfie does not get at least 10 documents matching your query request it will automatically move to the next three and so on until all are searched or until 10 matches are found.

If Arfie does find 10 or more matches you still can go to the next set of search engines by pressing on the button at the bottom of the page.

Engines have generally been placed in order from the very general index search (where a general search like "usenet culture" will not turn up 30000 pages) to the very specific super-engine (which will find too much if your search is not narrow). This means that you can put in as much or as little detail about what it is you want to find and not be disappointed with too few matches (since Arfie automatically fetches more documents from larger databases if less than ten are found) nor be overwhelmed when looking for information on a general topic (since the index search engines like Yahoo are categorized into general subject headings).

Attempts have been made to make it easy to follow up on the query sent to each search engine. The query sent to the search engine is printed and (in most cases) is linked to the page generating the results. Also if the search engine found more than the maximum displayable matches (e.g. 10 matches), a link to the next 10 should also be present (where supported).

Please remember that search engines change their format all the time. Thus Arfie is guaranteed NOT to work 100% of the time with 100% of the engines.

MetaCrawler is a World Wide Web search service developed in 1994 at the University of Washington by Erik Selberg, Oren Etzioni and Greg Lauckhart. It is now operated by go2net, Inc., an Internet content and technology company based in Seattle, Washington. In February 1997, go2net acquired the MetaCrawler from its original developers and is currently working on various improvements.

MetaCrawler differs from other search services in that it does not maintain any local database. Rather, it relies on the databases of various Web-based sources. *MetaCrawler sends your queries to several Web search engines, including Lycos, Infoseek, WebCrawler, Excite, AltaVista, and Yahoo.*

MetaCrawler queries the other search engines, organizes the results into a uniform format, ranks them by relevance, and returns them to the user. Of course, this means that MetaCrawler is slightly slower than other engines, but is more likely to obtain accurate results for your query.

Regular Search Vs Power Search   www.metacrawler.com/index_power.html

Power Search provides more options than a regular MetaCrawler search. Users are able to select results to be retrieved by continent/location, U.S. Educational Sites, U.S. Commercial Sites, and U.S. Governmental Sites. Results per page and a timeout option are included.

Consistent Searching Syntax

MetaCrawler offers a powerful search syntax, so you don't have to learn a different query language for each engine. In addition to the basic "any words", "all words", and "as a phrase" options, MetaCrawler recognizes a special search syntax that allows you to describe your desired results very specifically.

Service Vote Rankings

MetaCrawler combines and normalizes the confidence scores given to each reference by the services that return it. Thus, when MetaCrawler returns a reference, it sums the scores given by each service and presents them in a "voted" ordering, with the score (from 1 to 1000) presented in bold type next to each result.

Ask Jeeves allows you to ask a question in plain English and, after confirming the question, Ask Jeeves takes you to one and only one web site that answers your query.

A kids' version, Ask Jeeves for KidsTM, is located at AJKids.com and provides a safe and easy way for kids to find information on the Internet.

With Ask Jeeves, unlike other search engines, users enter a question in plain English and then Ask Jeeves presents a list of matched questions (typically, just a few). After the user selects the closest match, Ask Jeeves takes him or her directly to a site that was selected by the Ask Jeeves research staff as being an appropriate answer to the question. The user never faces a dreaded response such as "7837 matches to your query"!

For example, if you ask, "Who is the king of Siam?", Ask Jeeves would respond with "Who is the head of state of Thailand?". When you click on that question, Ask Jeeves takes you to a particular page on a site that presents information about the Thai king.

Is Ask Jeeves a metasearch?

No. Ask Jeeves uses its own knowledgebase to answer your question. However, because Jeeves doesn't know the answer to every question in the field of human knowledge (yet!), he also provides summarized results from passing your query on to several conventional search engines. This "metasearch" function is provided as a back up to Jeeves own question answering service, in the same sense that Yahoo provides an AltaVista keyword search function as a back up to its directory service.

How does Ask Jeeves work?

Ask Jeeves uses sophisticated natural language processing to understand and match users' questions to an extensive knowledgebase. The Ask Jeeves knowledgebase consists of thousands of question templates and millions of researched answer links to web sites. Examples of question templates are "Why is the sky blue?" and "Where can I find a map of [Name of City]?". In the first example, there is only one answer link matched to the question; in the second example, there are thousands (one for each city). The Ask Jeeves research staff selects questions and then searches the Internet for the best answer sites. This saves users countless hours of searching on their own. The Ask Jeeves knowledgebase is built by humans, not by software "spiders" and, therefore, each answer link is guaranteed to be relevant to the question asked.

Librarian's Index to the Internet

sunsite.berkeley.edu/InternetIndex/index.html

The Librarians' Index to the Internet is a searchable, annotated, subject directory of more than 3,000 Internet resources selected for their usefulness to the public library user's information needs.

The Index began in 1990 as Carole Leita's Gopher bookmark file. It migrated to the Berkeley Public Library's Web Server in 1993 as the Berkeley Public Library Index to the Internet. In late 1996, Carole began working with Roy Tennant at the Digital Library SunSITE to add a search engine to it (SWISH-Enhanced), add subject index terms, and create a system whereby other librarians would be able to add entries to the Index.

In March, 1997, the Berkeley Public Library Index to the Internet was moved to the Berkeley SunSITE and became the Librarians' Index to the Internet. The search engine is now in place, as are the subject terms.

Search Strategies

- ALL (the default) - searches all fields - title, subject, and annotation.
- Subject - searches the assigned subject field, loosely based on the Library of Congress Subject Headings. Use when you have a general topic in mind and don't see the subject in the categories list. You can browse our list of ALL subject terms used.
- Titles - searches the title field. Use when you know at least one keyword of the title of a resource.
- Annotations - searches just the description field of the resource.

A Boolean "and" between words is assumed; that is, documents will be retrieved that have all the specified words. If you wish to find all the documents that have any word, then use "or" between your search words.

Example: digital or virtual or electronic library

You can also use the Boolean operator "not" to eliminate words.

Example: censorship not filtering

To truncate a word, use an asterisk (*) as an operator at the end of the word. For example, the search "librar*" would retrieve documents that have the words "library", "libraries", "librarian", etc. Note: The new version of our search engine - SWISH-E (Simple Web Indexing System for Humans - Enhanced) now allows truncation to be used in combination with other search terms.

Example: filter* librar*

Citing Internet Resources

Many people want to know how to cite information that they find on the Internet in school papers, theses, reports, etc. There is no definitive answer, but many people have made suggestions.

The basic components of the reference citation are:

Author's Lastname, Author's Firstname. "Title of Document."
Title of Complete Work (if applicable). Version or File Number, if applicable.
Document date or date of last revision (if different from access date).
Protocol and address, access path or directories (date of access).

For example:

Burka, Lauren P. "A Hypertext History of Multi-User Dimensions."
The MUDdex. 1993.
www.apocalypse.org/pub/u/lpb/muddex/essay/ (5 Dec. 1994).

A recommended book for library collections is:
Electronic styles: a handbook for citing electronic information by Xia Li and Nancy B. Crane. Medford, NJ: Information Today, 1996 (2nd Edition) ISBN: 157387-027-7, $19.99.

Listed below are some places to go for recommended electronic information citation guides.

Internet Public Library (general information)
www.ipl.org/ref/QUE/FARQ/netciteFARQ.html

For citations for U.S. government publications in electronic format, see
www.lib.memphis.edu/gpo/citeweb.htm#online

The Spider's Apprentice

04/17/1997

How To Use Web Search Engines
Tips on using internet search sites like Google, alltheweb, and Yahoo.

Historical Search Engine Information -- Our In-Depth Analysis of Popular Search Engines, circa 1996-98

Important Note: This page is outdated. Several of these search engines no longer exist or are no longer used as much as they used to be. The rankings below are from the 1996-98 time period.

We are leaving the page up for historical purposes only -- some researchers are interested in the history of web search engines.

AltaVista

Alta Vista is a fast, powerful search engine with enough bells and whistles to do an extremely complex search, but first you have to master all its options. If you're serious about Web searching, however, mastering Alta Vista is a wise policy.

Type of search: Keyword

Search options: Simple or Advanced search, search refining.

Domains searched: Web, Usenet

Search refining: Boolean "AND," "OR" and "NOT," plus the proximal locator "NEAR." Allows wildcards and "backwards" searching (i.e., you can find all the other web sites that link to another page). You can decide how search terms should be weighed, and where in the document to look for them. Powerful search refining tools, and the more refining you do, the better your results are.

Relevance ranking: Ranks according to how many of your search terms a page contains, where in the document, and how close to one another the search terms are.

Results presented as: First several lines of document. "Detailed" summaries don't appear any more detailed than "standard" ones.

User interface: Reasonably good, but not very friendly to the casual user. Advanced query now allows you to further refine your search at the end of each results page. You can also visit specialized zones or channels in areas like finance, travel, news.

Help files: Complete, but confusing. Too much thrown at you at once. More clarity and more explanation of options would be appreciated!

Good points: Fast searches, capitalization and proper nouns recognized, largest database; finds things others don't. Alta Vista searches both the Web and Usenet. It will search on both words and on phrases, including names and titles. You can even search to discover how many people have linked their site to yours. You can also have the resulting pages of your searches translated into several other languages.

Bad points: Multiple pages from the same site show up too frequently; some curious relevancy rankings, especially on Simple search.

Overall Rating: A-

Excite

Spidap Tidbits--Did you know this? America Online made a deal with Excite, giving AOL a share in the company and making Excite AOL's partner and official search engine. In fact, AOL is now using the Excite engine as their proprietary AOL search engine, accessible both from within the AOL network and via the Web. Check out AOL's NetFind.

Excite bills itself as the "intelligent" search engine because of its concept-based indexing. While "intelligent" is an exaggeration (the apparent intelligence comes from the clever use of statistics, not from a sudden advance in artificial intelligence), Excite *is* one of our favorite search tools.

Type of search: Both concept and keyword

Search options: Simple, refined

Domains searched: Web, Usenet and classified ads

Search refining: Suggests you use more words, repeating key choices several times. Uses a fuzzy AND, which searches AND and OR, giving preference to AND. Has recently added Boolean operators to aid in search refining--AND, OR, AND NOT, and the characters + and -.

Relevance ranking: Confidence percentile provided on all searches, derivation unclear.

Results returned in: Summaries; will also sort them by site. By clicking on an icon beside each summary, you will get a cross-reference of similar sites.

User interface: *Generally good, nothing exciting.*

Help files: Very good, including a handbook that explains the site, the Web, the software, and how best to use their site.

Good points: *Large index. Not quite as up-to-date as it used to be. Excellent summaries, which they admit are actually highlights--the top few most important sentences in the document. You can view your hits in various ways, too--grouped by confidence or grouped by Web site.*

Bad points: Does not specify the format or the size in megabytes of the hits it returns, nor does it tell you upfront exactly how many hits there are.

Overall rating: B

Infoseek

Type of search: Keyword

Search options: Simple, but powerful (see comments below). Infoseek now uses the Ultraseek engine, which really zips along. The site has added an extensive catalogue section for subject-oriented searching. You can also cross-reference your search terms with similar catalogue subject items and searches come back with subjects automatically appended. You can also search images, which seems to be popular suddenly.

Domains searched: Web, Usenet, Usenet FAQs, Reviews, Topics.

Search refining: Phrases, capitalization, no Boolean operators, but uses + and - instead (similar to AND and NOT).

Relevance ranking: Gives numerical scores based on frequency and comparison to words already in their database.

Results presented as: First 30-100 words of the page

User interface: Good, easy to use, clear. Infoseek is also now allowing free searches of some of its extensive databases (stock quotes, company information, e-mail addresses, various reference works like dictionaries and zip code directories).

Help files: Good, useful.

Good points: Fast, flexible, reliable searching. Good output, which gives the URL, the size of the document and the relevancy score. Allows you to see similar pages (based on topic information about the pages). Full-text indexing, allows capital letters and phrases.

Bad points: We're sure Infoseek has some bad points, but we really can't think of any offhand!

Overall Rating: A-

Lycos

Type of search: Keyword, but Lycos is gradually becoming less of a search engine, it seems, and more of a Yahoo-like subject index. Has recently had a cool graphical facelift. Proud of its ability to search on image and sound files.

Search options: Basic orAdvanced

Domains searched: Web, Usenet, News, Stocks, Weather, Mult-media.

Search refining : Lycos now has full Boolean capabilities (using choices on drop-down forms).

Relevance ranking: Lycos no longer provides a relevancy ranking.

Results presented as: First 100 or so words in simple search, you choose in advanced search--summary, full results or short version.

User interface: Clean, clear, focuses more on directory now than on simple search.

Help files: Good, informative, graphical help screens are easy to understand.

Good points: Large database. Comprehensive results given--i.e., the date of the document, its size, etc. Lycos indexes the frequency with which documents are linked to by other documents to make sure the most popular web sites are found and indexed before the less popular ones.

Overall Rating: B+

Webcrawler

Spidap Tidbits--Did You Know This? AOL owns Webcrawler, but AOL's new deal with Excite means that the Webcrawler search engine and directory will be incorporated into Excite.

Type of search: Keyword

Search options: Simple, refined

Search options: Domains searched: Web, Usenet

Search refining : Uses either "and" or "any." Webcrawler has added full Boolean search term capability, including AND, OR, AND NOT, ADJ, (adjacent) and NEAR.

Relevance ranking: Yes--frequency calculated--computes the total number of times your keywords appear in the document and divides it by the total number of words in the document. Webcrawler returns surprisingly relevant results.

Results presented as: lists of hyperlinks or summaries, as the user chooses.

User interface: Good--easy and fun to use

Help files: Useful tips and FAQ.

Good points: Easy to use. Popular on the Web because it belongs to AOL and there are a lot of websurfers who sign on from AOL. Publishes usage statistics on their site. Also provides a service by which you can check to see whether a particular URL is in their index, and, if so, when it was last visited by their "spider." There is also some fascinating information about how Webcrawler's search strategy works.

Bad points: Speed seems to be slowing down a little recently. Its previous weakness--no way to refine search--has been eliminated with the addition of Boolean operators.

Overall Rating: B-

HotBot

Type of search: Keyword

Search options: Simple, Modified, Expert

Domains searched: Web

Search refining: Multiple types, including by phrase, person and Boolean-like choices in pull-down boxes. No proximal operators at present. In Expert searches you can search by date and even by different media types (Java, Javascript, Shockwave, VRML, etc.).

Relevance ranking: Yes. Methods used--search terms in the title will be ranked higher search terms in the text. Frequency also counts, and will result in higher rankings when search terms appears more frequently in short documents than when they appear frequently in very long documents. (This sounds sensible and useful).

Results presented as: Relevancy score and URL

User interface: Very cool and lively. Some users have complained about the bright green background, but we kinda like it.

Help files: A FAQ that answers users' questions, but not a lot of serious help files.

Good points: Claims to be fast because of the use of parallel processing, which distributes the load of queries as well as the database over several work stations.

Bad points: Some limitations still on Boolean operators, and the help files still aren't very good.

Overall Rating: B

Yahoo

Although not precisely a search engine site, Yahoo is an important Web resource. It works as an hierarchical subject index, allowing you to drill down from the general to the specific. Yahoo is an attempt to organize and catalogue the Web.

Yahoo also has search capabilities. You can search the Yahoo index (note: when you do this you are *not* searching the entire Web). If your query gets no hits in this manner, Yahoo offers you the option of searching the Alta Vista, which *does* search the entire Web.

Yahoo will also automatically feed your query into the other major search engine sites if you so desire. Thus, Yahoo has the capacity to act as a kind of meta-search engine.

Type of search: Keyword

Search options: Simple, Advanced

Domains searched: Yahoo's index, Usenet, E-mail addresses. Yahoo searches titles, URLs and the brief comments or descriptions of the Web sites Yahoo indexes.

Search refining: Boolean AND and OR. Yahoo is case insensitive.

Relevance ranking: Since Yahoo returns relatively few hits (it will never return more than 100), it's not clear how results are ranked.

Results presented as: Yahoo tells you the category where a hit is found, then

gives you a two-line description of the site.

User interface: Excellent, easy-to-use

Help files: Not very complete, but since there aren't a lot of search options, detailed help files are not necessary.

Good points: Easy-to-navigate subject catalogue. If you know what you want to find, Yahoo should be your first stop on the Web.

Bad points: Only a small portion of the Web has actually been catalogued by Yahoo.

Overall rating: A (This rating refers simply to Yahoo's quality as a directory--searches of the entire Web are not possible).

Spidap, Top Page

Contact Us

This Webliography / Bibliography was last modfied for the WI Educational Media Association Conference (4-24-97) program titled:

A Higher Signal - To - Noise Ratio:
Effective Use Of Web Search Engines

Webliography / Bibliography

Compiled by:

Bob Bocher, Library Technology Consultant
Kay Ihlenfeldt, Librarian, Master Searcher
WI Dept. of Public Instruction
WI Division for Libraries And Community Learning
bocherf@mail.state.wi.us - or- kay.ihlenfeldt@dpi.state.wi.us

There are many good, detailed resources available both on and off the Web to help you better understand how to use various search engines to best meet your needs. The list below represents a broad cross section of resources on Web search engines. Because this is a rapidly changing area most of the sites below are no more than a year old. About half of the citations on this page are from the page *Sink or Swim: Internet Search Tools & Techniques.* They appear here with permission of the author, Ross Tyner.

● Barlow, Linda. *The Spider's Apprentice: How To Use Web Search Engines.* April 17, 1997. Available at: http://www.monash.com/spidap.html

 • Very good site providing reviews of the popular engines listing good and bad points, a section on search strategy and a basic search engine FAQ.

● Birmingham, Judy. *Internet Search Engines*. March 13, 1996. Available at: http://www.stark.k12.oh.us/Docs/search/

 • Succinct listing in table format of features from the major search engines.

● Brandt, D. Scott. "Relevancy and Searching the Internet." *Computers in Libraries* 16.8 (September 1996): 35, 38-9.

● Campbell, Karen. *Understanding and Comparing Search Engines*. April 1996. Available at: http://www.hamline.edu/library/links/comparisons.html

 A Meta-list of 11 other sites that critique search engines.

● Campbell, Karen. *Tips on Popular Search Engines.*. March 1997. Available at: http://www.hamline.edu/library/bush/handouts/slahandout.html

 A good summary of several popular engines including Alta Vista, Excite, Lycos, InfoSeek, etc. Includes a table of comparing various features.

What are "Meta-search" Engines?

In ordinary search engines or search tools (such as Infoseek, AltaVista, Yahoo!, Hotbot, or Excite), you submit keywords to a single database of web-pages owned by the search tool, and you get back a different display of documents from each search engine's unique database of web-pages. Results from submitting very comparable searches can differ widely, but also contain some of the same sites.

In a meta-search engine, you submit keywords in its search box, and it transmits your search simultaneously to most of the popular search engines and their databases of web pages. Within a few seconds, you get back a compilation of results containing matching sites from all of the search engines queried. This can save you a lot of time and provide an overview of the kinds of documents "out there" matching any term, phrase-in-quotes, or set of terms and phrases.

Meta-search engines do not own any database of web-pages; they use and deliver the databases and searching programs of each of the popular, individual search tools they query. Meta-search engines act as intelligent middle-agents to pass your search through, gather the responses from the individual search tools they query, and then give you a more unified report of results from many different resources.

How do I choose which one to use?

All of the meta-search engines listed here produce very adequate search results, and have certain features in common:

Dogpile   www.dogpile.com
Inference Find   www.inference.com/infind
MetaCrawler   www.metacrawler.com
Metafind   www.metafind.com
AskJeeves   www.aj.com

- They all search most of the popular search engines. All search most of the search tools this tutorial recommends.
- They are all fast, because they use "parallel" (i.e., simultaneous) querying of the individual search tools and have high-speed processors to format and deliver the results to your screen.
- They all allow you to set the length of time you are willing to wait and to personalize some aspects of the format. The longer the time, the more results you will get.

What are "Meta-search" Engines? - Continued

They differ in these other significant features:

- How results are compiled when reported. Some report the results from each search engine in sequence, giving you a list from each in order queried. Others sort the results, eliminating duplicates. In some you can specify how results are sorted; in others the default is significant phrases or words.

- How and whether they can handle complex searches. Some allow phrase searching, some allow Boolean operators (especially OR and NOT) for the search tools that support Boolean operators. Some strip out quotations or Boolean operators, or create garbage by passing them through as search terms. Few allow you to request truncation. In some you have more flexibility to vary time limits and choose how results are reported. Some let you specify which search tool databases are queried and in what order.

For more information go to
www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html

What is a Metasite?

Metasites are comprehensive in coverage of one or more subjects.

for instance, "**meta**" is a prefix used:

[metaphysics] : more comprehensive : transcending <metapsychology> -- used with the name of a discipline to designate a new but related discipline designed to deal critically with the original one <metamathematics> © 1997 by Merriam-Webster, Incorporated

For instance, the South Carolina Reference Room would be considered a "**metasite**" for South Carolina related information.

URL: www.state.sc.us/scsl/refdesk.html

Other such web metasites are:

GPO Access	www.access.gpo.gov/su_docs/dbsearch.html
HealthGate	www.healthgate.com
FindLaw	www.findlaw.com
Librarian's Index to the Internet	sunsite.berkeley.edu/InternetIndex/index.html

WebSerch - The Web Research Resourc

W ebSerch

Extending your reach

Evaluate Search Engines

'Relevancy ranking algorithms often fail to lift the most relevant hits to the top' Notess, G. Online, Jul/Aug 1997, p.66

Requires critical thinking.

Discriminate between references.

'QRAQ'- quantity, relevance, authority, quality. (<u>1</u>)

- Response time. How quickly are results returned/displayed?
- Results display. What level of customization is available re. the range of output options (detailed, brief, title only), the number of retrieved records displayed per page?
- What does the citation indicate? Does it indicate the source, file size, URL, etc.? Does the citation indicate a date? If so, what does the date denote?
  Note -the date display on results page may refer to (a) the publication date, (b) creation date given in the page's HTML, (c) last modified date, (d) if none of the others is available, the date the search engine found the page.
- How useful is the site description? Does it consist of an abstract, extract? Does it display a user-provided description (from the meta tag descriptor field) where so provided?
  Note - all search engines provide some textual description of retrieved sites. Some use first N characters of document (which may often be meaningless information). Some use the HTML 'Meta' tag.
  Questions – How useful is the description in assessing relevance of retrieval? Is the description enough to allow the user to make a decision about whether or not to display the document?
- Do the results as displayed help indicate the degree of success of the relevancy ranking criteria as used by the search engine?
- Is the default relevancy ranking criteria used clearly stated?
- Where different areas of the database (eg. web, special collections, directory) are searched simultaneously, can results be sorted by such areas?
- If the search engine returns results from the different areas of its database separately, does relevancy ranking suffer?
- Can and does the search engine remove duplicates from the search results *(slightly different URL, same content)*? How does it treat mirror sites?
- Post-search processing facilities. What facilities are there for the sorting of results? Can they be sorted by site, by date?
- Can searches be saved and rerun?

What ability is there to move through the results display?

- KWIC search terms – are they highlighted in found documents?
- To what extent are there inactive links? What conclusions can be drawn from the level of inactive links? Does the reason given (eg. path/file not found) maybe suggest that the index is not updated frequently enough, or not at all? Page moved, server not responding or down - what conclusions can be drawn from same?

- **Recall** *(degree to which search engine returns all the matching documents in a collection - requires knowledge of the total number of matching/relevant documents in a collection. The higher the recall, the more efficient the search.).*
- **Coverage** *(ratio of matching documents found as opposed to the total number of matching documents engine could have found).*
- **Precision/accuracy** *(degree to which search engine lists documents matching a query, or, in other words, the fraction of the search output relevant to a particular query).* Points to the search engine's ability in querying its index and returning relevant pages whilst filtering out irrelevant pages. From this can be adjudged how 'smart' the search engine technology is. The one variable is the searcher's ability to be effective in structuring the query.
- **Relevance** *(how well document matches users request - determined by assessing the subject matter).* (1)

  How relevant is subject matter? (2)

  o <u>Irrelevant links</u>. Does not satisfy important aspect of the search expression.
  o <u>Technically relevant</u>. Page satisfies query but is not potentially useful (not related to topic indicated or too small or uninformative).
  o <u>Potentially useful.</u> Some useful information about the correct topic - could be of some conceivable use to some searcher - also provides links to pages adjudged to be most probably useful.
  o <u>Most probably useful</u>. Useful to almost anyone who would conduct the search. Bibliography (webliography) - points to pages that deal with many aspects of the topic - page is adjudged to be extremely thorough for the topic.

Note: Mirror sites are not duplicate sites - different URL, slightly or somewhat different content.

- **Reliability SEE** <u>Evaluation of Web Resources</u>

References

(1) For in-depth discussion, see Clarke, S. J. and Willett, P., 1997. *Estimating the recall performance of Web search engines.* Aslib Proceedings, Vol. 49, no. 7 July/August '97, pp.184-189

also Duff, A., 1996. *The Literature search: a library-based model for information skills instruction.* Library Review, Vol. 45, no. 4, 1996, pp.14-18.

(2) see Leighton, H. V. & Srivastava, Dr. Jaideep., 1997. *Precision among WWW search services (search engines): AltaVista, Excite, HotBot, Infoseek*

Contents
Introduction
Index Composition
Search Capability
Retrieval Perf rmance
Usability

| Home Page | ▾ | Go |

View Metadata for this page!

Back to top

Launched: Nov'99
Update:25Jun'00

© 1999-2000
Eddie Byrne

Launched: Nov'99

Link to this
site

Doras Review

Pandia
Review

Contains

Dublin
Core

Metadata

Contact the author at:
edwardbyrne@ireland.com

Copyright notice

URL: **www.clubi.ie/webserch/**
also **www.webserch.com**

Site design: Eddie Byrne. Contact at webmaster@webserch.com

EasySEARCH

1996, 1997, 1998, 1999

Monthly hit score of each search engine

engines	1996									
	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	TOTAL
a2z	-	-	1300	677	520	460	465	562	573	4557
Alta Vista	1040	4345	8922	12328	12925	14825	15167	15222	14208	98982
Casbah	-	-	-	-	1079	1003	1002	1022	858	4964
Deja News	-	77	391	558	520	479	529	524	701	3779
Ecila	-	-	-	-	-	1115	752	699	577	3143
Excite	222	1087	1751	1812	2101	1908	1805	2395	1787	14868
HotBot	-	-	2199	1211	957	897	925	1157	1125	8471
Infoseek Guide	1197	4985	5251	4585	3938	2031	1374	1287	1382	26030
Infoseek Ultra	-	-	-	-	-	4443	4030	4148	3543	16164
Inktomi	221	1131	1423	1080	998	816	714	672	516	7571
Link Star	-	62	536	412	495	470	469	855	744	4043
Lycos	473	1800	2594	2742	3007	2960	2927	2765	2540	21808
Magellan	188	709	965	840	982	748	463	432	457	5784
Meta Crawler	24	112	493	414	424	436	502	373	393	3171
OpenText Index	185	598	935	810	960	795	681	654	655	6273
Point	-	464	505	442	445	810	476	421	399	3962
Robo Link	-	-	-	-	-	-	-	1811	1300	3111
shareware com	122	552	692	594	638	546	503	525	560	4732
Wab Crawler	236	1022	1468	1389	1581	1419	1149	1240	1200	10704
Who Where	165	803	1052	933	848	787	702	739	651	6680
Yahoo!	541	2645	4934	5461	7129	8549	8067	9138	8219	54683
ZD Net	-	67	641	559	634	688	614	715	587	4505
Acara (Webdew)	-	-	2093	52	3296	3576	3533	3387	3888	19825
Asahi DNA	-	-	2998	3330	3932	5461	5814	3329	535	25399
CSJ Index	586	4224	6557	6493	6893	7354	6435	5845	5912	50299
Dragon (Japan S.E.)	625	4263	5403	5218	5314	4835	3498	3742	4182	37080
Hole in One	781	6108	6927	7056	7294	6513	5290	5012	5086	50067
Info Navigator	-	384	2966	2887	3783	4097	3989	4415	4597	27118
Infoseek JAPAN	-	-	-	-	-	-	4799	12225	13074	30098
Mondou	685	3995	5406	5019	5291	5479	4918	4885	4992	40670
Net Plaza	-	493	2658	2907	3306	4352	3761	3936	4136	25549
Nippon S.E.	-	478	3764	3539	3818	4001	3687	3816	4209	27312
NTT Directory	-	552	2530	3060	3364	4072	3783	3943	4390	25694
ODIN	1130	6368	11419	12533	12635	14611	13337	11257	10256	93546

JAPAN		2368	11419	12533	12635	14611				
Snrigan	1605	10757	16675	16163	17541	22188	20014	23452	24357	152752
TITAN	793	5059	7946	8946	9672	10381	9234	8713	9197	69941
Yahoo! JAPAN	1077	8394	14364	16451	19206	21298	17728	19145	19654	137317
TOTAL	11896	71534	127758	130501	145526	164403	153136	164458	161440	1130652

honda@nibh.go.jp / National Institute of Bioscience and Human Technology

EasySEARCH

1996, 1997, 1998, 1999

Monthly hit score of each search engine

engines	1996 TOTAL	1997 Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oc
a2z	4557	496	-	-	-	-	-	-	-	-	
Alta Vista	98982	15906	8647	5182	3570	3988	3843	3108	3220	3391	38
Casbah	4964	792	534	273	167	141	173	100	131	118	1
Deja News	3779	794	473	266	166	164	159	118	152	153	1
Ecila	3143	582	412	305	270	293	307	294	317	369	3
EuroSeek	-	-	-	1181	704	747	673	622	700	632	6
Excite	14868	2135	1338	870	663	541	559	554	628	636	6
Four11	-	-	60	155	83	96	90	68	60	81	1
HotBot	8471	1528	1068	749	609	739	788	642	828	943	9
ImageSurfer	-	-	116	512	357	437	317	334	343	361	3
Infoseek Guide	26030	675	14	3	-	-	-	-	-	-	
Infoseek Ultra	16164	4313	2772	1663	1297	1412	1281	1263	1396	1554	15
Inktomi	7571	-	-	-	-	-	-	-	-	-	
Link Star	4043	599	367	190	219	206	214	177	221	234	2
Lycos	21808	2707	1623	996	671	616	477	447	430	430	3
Magellan	5784	560	352	269	235	193	244	186	190	238	2
Meta Crawler	3171	367	385	185	181	115	197	183	172	197	2
OpenText Index	6273	719	519	299	257	226	190	147	160	189	1
Point	3962	449	297	212	178	146	196	168	143	176	1
Robo Link	3111	1401	616	461	343	300	248	213	190	-	
shareware com	4732	544	341	180	163	133	132	92	88	90	
Wab Crawler	10704	1131	695	403	251	250	273	270	273	297	4
Who Where	6680	562	363	218	194	219	183	135	168	161	1
Yahoo!	54683	8680	4676	2581	2036	2154	2324	1640	1654	1865	19
ZD Net	4505	523	423	194	147	126	87	87	90	97	1
Acara (Webdew)	19825	4609	6585	6264	4572	5799	4789	4623	4653	4341	48
Asahi DNA	25399	486	126	-	-	-	-	-	-	-	
Books	-	-	-	-	-	-	-	-	-	197	50
CSJ Index	50299	6164	7551	6599	4479	4571	3811	3742	3268	4908	

Search Engine											
Dragon (Japan S.E.)	37080	4844	5154	3772	2825	2976	2450	2547	3323	3586	43
ExciteJ	-	-	-	-	-	-	-	1929	17174	15284	111
Goo	-	-	-	7506	20045	49151	48838	47038	45089	54797	662
Hole in One	50067	5743	7175	4898	2994	3348	3168	2960	3115	2815	28
ImageSurferJ	-	-	1324	3024	1493	1598	1388	1488	1584	1929	19
Info Navigator	27118	4721	5911	5695	3581	3532	2603	2378	7614	6120	61
Infoseek JAPAN	30098	13738	14606	13628	10287	14103	12910	15538	14279	16458	244
JOY	-	-	-	4935	2191	2200	2027	2098	2075	2096	23
Mondou	40670	4964	5635	3830	2285	2440	2049	2001	1969	1938	22
Music.co.jp	-	-	-	539	1242	1575	1255	1100	1274	1301	15
Net Plaza	25549	4435	5016	3912	2738	3174	2720	2841	6626	5494	50
Nikkei	-	-	-	4452	2626	2750	2360	2107	2508	2605	31
Nippon S.E.	27312	4809	4869	3290	1926	2143	1570	1563	1657	1709	18
NTT Directory	25694	4323	5576	3743	2409	2751	3500	2763	2341	2283	29
ODIN	93546	23412	17677	15290	9692	7596	6252	5429	3719	3134	32
OpenText IndexJ	-	-	2060	3727	2046	2409	1982	2019	1930	1902	22
Snrigan	152752	26750	31151	27203	26992	24458	17149	14983	11643	11945	146
TITAN	69941	8885	10403	9411	6451	6756	5533	6107	5788	4938	55
VectorPack	-	-	-	549	1256	1605	1463	1270	1421	1572	19
Yahoo! JAPAN	137317	22994	24388	25177	22855	24375	20039	19344	17633	22445	289
Yubinya	-	-	-	802	1885	1889	1601	1423	1667	1847	19
TOTAL	1130652	186340	181298	175593	149631	184441	162412	158139	173904	187856	2278

honda@nibh.go.jp / National Institute of Bioscience and Human Technology

Submit Corner

Wordtracker - discover the keywords [⊟][□][⊠]

`<html><meta name="keywords" content="want to know the best keyw rds for your meta tags ?">|` [Click Here]

Submit Corner : Guides : Search Engine Guide : Metacrawler

Toolshop

Assessment Tools
Keyword Thesaurus
Link Popularity Tracker
META Tag Generator
META Tag Scanner
Robot Generator
Submit Engine
Top Keywords NEW!
Word Tracker NEW!

Guides

Bandwidth Conservation
Background Images
External Javascript
Optimize HTML
See Complete List

META Tags
Description Tag
Keywords Tag
Revisit Tag
See Complete List

Search Engines
Altavista
Excite
Yahoo!
Stability Report
See Complete List

Site Improvement
Image Tags
Title Optimization
Effective META Tags
See Complete List

Software Packages NEW!
Web Position Gold
Submit Wolf
See C mplete List

W bsite Assessment
Editor's Viewpoint of
Directories NEW!
Assess Your Site Indexing
Who's Linking To You
Competitor Identification

FAQ

Metacrawler

By Submit Corner
Tell a Friend About This Page

Overview: Metacrawler is a meta search engine which queries multiple sites at once including major industry players

Originally launched in 1995 by a graduate student and an associate professor from the University of Washington, Metacrawler became a popular meta search engine. In February of 1997, Go2Net acquired Metacrawler and now owns one of its competitors, Dogpile. Metacrawler has been voted twice by PC Magazine as Best Search Engine and has won other various awards. Today, Go2Net owns two of the largest meta search engines on the web and together queries the top search engines including Altavista, Looksmart, Lycos, GoTo, Direct Hit, Google, and Infoseek (Go Network).

Website Information

Metacrawler http://www.metacrawler.com
(New window opens)

Submission & Ranking Process

Metacrawler Characteristics	A meta search engine itself does not index or rank websites. Instead, it queries multiple websites simultaneously and will deliver a combined results page of the best sites from each search engine. The higher you rank on each individual search engine, the higher your site will rank on a

Sponsor Message
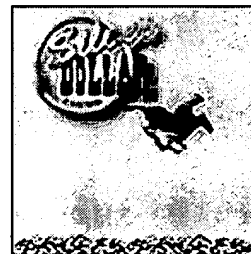
Consulting Services

Not sure how to implement the strategies we talk about here? Take all the hassles out and let us do it for you with our consulting services. See what we can do for your business.

Win Great Prizes Just for Using Our Services

Latest Headlines

All links open in a new window
View All Headlines
(Thu, Jan 29 19:00:01)

Ask Jeeves profit triples in fourth quarter
Source: San Jose Mercury News
Date: Jan 29 2004 11:17PM

Keep your Google mania in check
Source: Canadian Press via Canada.com
Date: Jan 29 2004 10:34PM

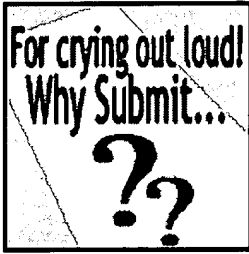Paid Search Bolsters Search Engine Revenue
Source: Search Engine Lowdown
Date: Jan 29 2004 10:20PM

Google Still Under Fire Over Trademark Dispute
Source: Search Engine Lowdown
Date: Jan 29 2004 10:20PM

Our Recommendations
Read FAQ

Inside Submit Corner

Advertise
Author Guidelines
Contact Us
Linking Info
Join Our Mailing List
Privacy Policy
Rave Reviews
Terms of Use

Ordering R sults

Metacrawler sorts its results based on the number of times a website was found in the top rankings. For example, if your site was found in the top 10 on 5 search engines, your site will be ranked above a site who was found in the top 10 on only 2 search engines. The higher your site ranks on a search engine, the better chances you have on getting higher in Metacrawler.

Getting Listed (Comments)

Metacrawler does not accept submissions directly through their site. Rather, you must submit to any one of the search engine partners that they use in order to get listed. We attempt to cover as many search partners as possible within this guide. Use the links in the overview section of this page to locate a partner for additional information on how to optimize your rankings for each individual search partner.

Submission Time

Getting listed in Metacrawler is usually instantaneous once your site becomes ranked in a partner's search database (Wait time varies per partner). Since GoTo is the fastest submission partner, we suggest you use GoTo to submit your site through (see additional information on GoTo).

Related Links

Ask Jeeves Posts Strong Quarterly Numbers
Source: Search Engine Lowdown
Date: Jan 29 2004 10:20PM

Google Likely to Use Email for Advertising
Source: Search Engine Lowdown
Date: Jan 29 2004 10:20PM

Google May Delay IPO
Source: Search Engine Lowdown
Date: Jan 29 2004 10:20PM

Google dashes hopes of float
Source: news.com.au
Date: Jan 29 2004 10:17PM

There are 22 additional news headlines. **Click to View All Headlines**

| **Automated Submissi n Tool** | Use our <u>Submission Tool</u> to submit your site to all the major search engines |
| **META Tag Generat r** | Create better META Tags with our <u>META Tag Generator</u> |

<u>Privacy</u> | <u>Terms of Use</u>

W ebSerch

Extending your reach

WebSerch - Th W b Research Resourc

Home	What's New	Contact	Site Map	About WebSerch
Search Tips	FAQs	Links	Glossary	Further Reading

Glossary

Acrobat	*see* PDF
ActiveX	A language which provides for dynamic content on a web page in the form of animations, video sequences and virtual reality displays. ActiveX controls include Shockwave and RealAudio. A technology developed by Microsoft.
ALT text	Alternative text - text placed within an image tag which will display instead of an image where the browser is unable to display images, or where the user may turn off image display to facilitate faster loading of pages. Also recommended as an accessibility feature for screen readers.
Anonymous Login	The means of logging on to an FTP site. 'Anonymous' is given as the user name, while the user's e-mail address is given as the password
Applet	A small program embedded in a web page which can perform a particular task. *See also* Java.
Archie	A tool which indexes FTP sites, and then allows you to query that index. To use Archie, one needs to access a server (via the Telnet protocol for connecting to a remote computer) which hosts Archie and enter its particular commands.
ASP	Active Server Page - a page containing script(s) which is processed on a Microsoft Internet Information Server before being sent back to the user. In other word, a page created on the fly, usually as a result of the user inputting details on a form requesting information. The returned page is therefore a customized response based on a user's initial request. Is an alternative to CGI.
Blind link	A link which doesn't lead anywhere.
Boolean operators	The logical operators AND, OR, and NOT. AND indicates terms must be included, NOT indicates a term must be excluded, and OR indicates either term can be available in a retrieved document. Combining of terms is done by using parenthesis.

Cache	In web parlance, an area of memory set aside for storage of previously downloaded pages. PCs and proxy servers make use of cache memory for quicker retrieval of material. They first access the cache when a request is received. Providing the requested page from the cache if so available allows for a quicker response to the user's request without having to access the source server for the information. (If uncertain as to the recency of a retrieved page, holding down the Control button while hitting the browsers Refresh button will request the page from the host server). For an intersting perspective on cache, see Ivan Trundle's article in *InCite*, October 1999 issue, available online from: http://www.alia.org.au/incite/1999/10/cache.html [accessed 22 November 1999]
CGI	Common Gateway Interface - a method for passing information back and forth between the user's computer and the web server when the user seeks information by filling in a form. This is necessary when an element of processing is required giving a customized response to the user. Is an alternative to ASP.
Cookie	An ASCII text file downloaded onto a computer's hard drive by a site which then retains information about the user for when they next access the remote site. The profile kept on the user's own computer is intended to assist the site in identifying the user's preferences and as such is quite harmless. Other common uses include shopping baskets and interest profiles, which help a site customize advertisement displays to match a users browsing habits. Some sites require the user to accept cookies before allowing them access. However the user can configure their computer to refuse cookies, but this may make some sites inaccessible.
Coverage	That portion or quantity of the web visited and indexed by web crawlers.
Crawler	*See* Robot
Discussion group	*See* Newsgroups
Dead links	Links to pages that no longer exist.
False drops	Irrelevant returns
Field searching	The limiting of a search to a specific location as identified by the field identifier, e.g.. Title, URL, Domain, Link. The identifier is usually followed by a colon, then the word, with no spacing.

Frame	A means of opening more than one file (HTML page) at a time in the same window. A typical use of multiple frames in the same window is to present a menu in one frame, site logo in another, and content pages in the 'main' frame. This way, the menu and site heading do not need to be replicated on each page and retain the same position and consistency, and remain visible regardless of scrolling in another frame. Activating a link in one frame generally opens a page in another frame.
FTP	File Transfer Protocol - a protocol for the exchange of files between computers on the Internet. Used to download files or upload web pages to a server.
Fuzzy 'and'	The search engine will firstly list documents containing all the terms, then some of the terms until it finally lists documents containing only one of the search terms.
GIF	Graphics Interchange Format - an image format designed specifically for electronic transmission and ideally suited to the creation of simple web graphics like icons, logos and buttons, where the number of colours is limited. Allows for compression of file size with little loss in quality. The maximum number of colours that can be used in a GIF image is 256. It is generally not suitable for photographs because of this colour restriction. You can also have animated GIFs and transparent GIFs.
Gopher	A menu-based system which preceded the world wide web yet still exists, though largely superseded by the web. Files are kept on Gopher servers and can be accessed via web browsers. Usually referred to as Gopherspace, the primary search tool for searching Gopher file systems is Veronica.
Helper application	Software that is called upon to process file formats that the browser cannot. Examples of such formats include multimedia files and PDF files. Examples of helper applications include Realplayer and Acrobat Reader. Unlike plug-ins, there is no integration with the browser, though they perform similar functions. *See also* Plug-in
Home Page	The entry page of a web site which is the access point to the rest of the sites contents. Somewhat equivalent to the title and contents information of a book all being on the one page.
Homonym	Word with more than one meaning.
Hotspot	An area of an image which acts as a hyperlink to another location.
Hyperlink	The ability to link from one location to another by clicking on a word or graphic thereby activating a hyperlink. This ability is the foundation upon which the world wide web is based. The term

'hypertext link' is a narrower term which was more widely used when text was the main component of the web.

Image Map	A graphic image which has specific areas identified so that when clicking within that area, one activates a hyperlink and is taken to another destination. The areas within the image are identified by means of their pixel coordinates and then a URL is assigned to those coordinates.
Interlaced Image	An image which appears all at once instead of one line at a time as for a non-interlaced image. From initially appearing blurred, the image gradually sharpens until the whole image is downloaded. Therefore, all the image is visible sooner.
Invisible Web	That portion of the web which is not indexed by the search engines. The invisible web includes e.g. PDF files, dynamically generated pages, information included in databases, information protected by firewalls and password-protected sites. Search engines can also only index HTML documents to which they are given access.
IP Address	The four sets of numbers separated by periods which make up an Internet address. The Domain Name System (DNS) converts this numbering system into the more easily recognized URLs.
IRC	Internet Relay Chat.
Java	A programming language used to create interactive and animated web pages. The program is called an applet and is stored on the web server. It is downloaded onto the user's computer when a page containing the applet is accessed.
Java-enabled	Means a browser is capable of interpreting Java programming language.
Javascript	A script language created by Netscape allowing various actions to occur on a web page either automatically or as a result of a user's actions. Such actions may include a pop-up menu appearing, or a graphic changing during a mouse rollover. Though most browsers support javascript, they may interpret it in slightly different ways. Not to be confused with Java Javascript instructions are placed in the page's HTML.
JPEG	Joint Photographic Experts Group - an image format ideally suited to the display of photographic images and images requiring more colours that offered by the restrictive 256 colour palette. Can compress file size even more than the GIF format.

Listserv

A mailing list server which automatically forwards e-mail to everyone on a particular mailing list. One *subscribes* to a discussion group in order to participate and receive e-mail.

Meta-search engine

A search engine that submits your query to other search engines, thereby accessing several different databases simultaneously. It does not maintain its own database, and, depending on the sophistication of the particular meta-search engine, will often adopt a lowest common denominator approach in querying the engines covered. This means that often you cannot access the advanced features of the particular search engines queried by the meta-search engine.

Meta Tag

Tags included in the HEAD element of a web page and which, among other things, include descriptive information of the document. Information in the various Meta tags does not appear visible on the web document and is only viewable by looking at the page source. Web search engines often index the content of meta tags, thus allowing the page creator some opportunity to allocate suitable keywords. Search engines may also display the content of the Descriptor meta tag as part of their results display.

Metadata

Information about information. In the context of a web page, information enclosed in the meta tags included in the HEAD element of a document. *See* Meta Tag.

Mirror site

A close if not an exact replica of a main site, the purpose of which is to facilitate heavy traffic by spreading it among different servers and allowing for the fastest access possible. Some servers may have connections which allow for faster access to the web, and host sites may find that strategically placed servers (in a geographic sense) may facilitate faster access and downloads.

Moderated newsgroups

A newsgroups monitored by an authorized individual who can prevent messages from being posted if (s)he deems them inappropriate. The opposite is an *unmoderated* newsgroups.

MP3

MPEG-1 Audio Layer-3. A format for compressing sound into the smallest possible files while retaining sound quality. May require the downloading of an MP3 player if your browser does not have a player built-in.

Newsgroup

A discussion group about a particular topic to which subscribers can contribute by responding to previous postings (on bulletin boards) or create new topics. Non-subscribers can view postings but not contribute. Newsgroups are divided into different categories (e.g. rec, soc, comp), while most are unmoderated. Usenet is the network of newsgroups including the host computers and users. There are currently over 20,000 newsgroups.

OPAC	Online Public Access Catalogue - a library catalogue available via the Internet.
PDF	Portable Document Format - a file format created by Adobe which preserves the original appearance of a document. Ideally suited to reproducing and distributing exact copies of journal or magazine articles, and brochures. The files are created using Adobe's Acrobat product and are viewable using the Acrobat reader, which is a free download.
PNG	Portable Network Graphics. A newer bit-mapped graphics format similar to GIF and earmarked to replace it. It is a patent-free format supported by the latest browsers (IE4+, NN4+, Opera 3.6).
Plug-in	A program that can be downloaded and installed as part of the browser in order that certain file formats can be viewed, played or accessed in some way. If the browser does not have the appropriate plug-in to view a file, a prompt is given to download the relevant plug-in. Unlike helper applications, there is full integration with the browser. *See also* Helper Application
Portal	A term used for a site that proposes to be an entry point to the world wide web. Usually provides a host of services, customization features and interactive elements, as well as the traditional search features and directory services, the intention of which is to retain the visitor rather than merely forward them to another site with less chance of their return.
Precision	Degree to which search engine lists documents matching a query, or fraction of search output relevant to a particular query.
Proxy Server	A server that sits between the individual accessing the Internet. and the web servers hosting the information. Requests for information go through the proxy server which may try to satisfy the request from its own cache if it has the page held from a previous access. This gives quicker access to web information. Most Internet. Service Providers use proxy servers.
Quicktime	A technology which allows for the production of video and multimedia. Viewable with the Quicktime player.
RealAudio	A sound delivery technology which allows for sound to be heard as soon as the sound file starts loading, doing away with the wait period while the total file is downloaded. This is called *streaming sound*, and was developed by Progressive Networks.

Recall	Degree to which a search engine returns all the matching documents in a collection - requires knowledge of the total number of matching/relevant documents in a collection. The higher the recall, the more efficient the search.
Relevance	How well a document matches a user's request - determined by assessing the subject matter.
Robot	A program which accesses web pages by means of hyperlinks and sends copies of those pages back to a search engine for indexing. Will also visit submitted sites. Depending on the particular robot and which search engine it is from, it may only crawl a site to a particular depth and not all pages at a site. Some robots lay claim to knowing how often they need to revisit a site based on the site's frequency of revision. Most obey the robot exclusion protocol, and there are particular areas they cannot access (e.g. PDF files, firewall-protected sites). Also known as bots, spiders and crawlers.
Robot Exclusion Protocol	A protocol which allows for a web site creator to prevent search engines from accessing and indexing a site or a particular part of a site. This protocol is adhered to my most search engines. Enforced by means of a robot.txt file which is a file in the root directory of a web server.
Robot.txt	*See* Robot Exclusion Protocol
Spamming	The repeated use of keywords in order to boost the likelihood of a page being returned following a search. Is a technique frowned upon by search engines who supposedly can detect the technique and refuse to index the guilty site. The keywords used in spamming may have no relevance to the actual content, an irritation to the searcher when totally irrelevant material is returned.
Spider	*See* Robot
Stemming	Gerund, suffix stripping. This means a word is stripped back to a particular point or stem, and then searched on that stem plus common endings. The effect is to retrieve word variants, but may also retrieve unrelated words which share a common stem.
Stopword list	A list of common words which the search engines will not index because of their commonality, e.g. *the, or, an, a, if*. The stopword lists of the varied search engines may differ slightly. In some instances, search engines may add common word like *web* to its stopword list.
Subject Directory	A hierarchially-arranged tool using a self-styled classification scheme which is manually compiled and maintained. Pages included are usually reviewed and given some sort of rating. A subject directory

usually presents the choice of either browsing through its hierarchical scheme or querying the database with its own search engine. The emphasis with subject directories is usually on quality rather than quantity.

Synonyms	Different words with the same meaning.
Telnet	A means of accessing a remote computer. To access the remote computer, one needs permission and a userid to log on as though one was using the remote computer directly. Commands used after logging on are those available on the remote computer, not the user's own.
Term weighting	Putting most important term(s) first.
Thumbnail Image	A smaller copy of a larger image allowing the user to view the image and decide if they want to download the larger image. Useful in so far as it gives the user the choice where the larger image may take some time to download and may prove annoying in slowing up the downloading of a page. Sometimes the size of the larger image will be indicated (in kb) allowing the user to decide whether or not to download based on the size of the larger image.
Truncation	A means of finding word variants, usually by indicating a certain number of letters followed by an asterix (or similar operator). Can also, with certain search engines, do middle or beginning truncation as well as end truncation. Search engines usually require a minimum number of letters to be present in order to carry out a truncated search, and the operator (i.e. asterix) will usually replace any number of letters. However, the use of certain operators with particular search engines may replace only one letter if so programmed. Also known as a *Wildcard* search.
Unmoderated Newsgroups	*See* Moderated Newsgroups
URL	Uniform Resource Locator. A web page's address. Consists of protocol or file type indicator (http://, ftp://, telnet://, mailto:, gopher://, NNTP://), domain name (www.clubi.ie/), path indicating directory structure (webserch/engines/altavist/) and file name plus .htm or .html extension (display.htm). The domain name element consists of the top-level domain (www), second-level domain (company, organization name) and type and/or country domain indicator (e.g. .com, .edu, .co.uk, .ie).
Usenet	A network of newsgroups, host computers which use the Network News Transfer Protocol (NNTP) and the user population. *See also* Newsgroup.

Veronica   A spider which indexes Gopher servers and allows a user to query its database of file names. One needs to access a Gopher server which has Veronica on it via the Telnet protocol or via a web browser.

Wildcard   *See* Truncation

XML   Extensible Markup Language

Cannot locate the term? Search the TechEncyclopedia.

| Home Page | ▼ | Go |

View Metadata for this page!

Back to top

Launched: Nov'99   Link to this
Update:25Jun'00   site

                         Doras Review   Contains
© 1999-2000   Dublin
Eddie Byrne   Pandia   Core   Contact the author at:
                         Review   Metadata   edwardbyrne@ireland.com

Copyright notice   URL: **www.clubi.ie/webserch/**
                                    also **www.webserch.com**

                    Site design: Eddie Byrne. Contact at webmaster@webserch.com

Freeform Search

Database:
US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term:

Display: `10` Documents in **Display Format:** `-` **Starting with Number** `1`

Generate: ○ Hit List ◉ Hit Count ○ Side by Side ○ Image

[Search] [Clear] [Interrupt]

Search History

DATE: Friday, January 30, 2004   Printable Copy   Create Case

<u>Set Name</u>	<u>Query</u>	<u>Hit Count</u>	<u>Set Name</u>
side by side			result set
	DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR		
L15	L13 and (estimate or rank or calculate)	510	L15
L14	L13 and (estimate or rank or calculate) near (coverage or cover)	0	L14
L13	L12 and meta	964	L13
L12	search near engines	8985	L12
	DB=USPT; PLUR=YES; OP=OR		
L11	5867799.pn.	1	L11
L10	6012053.pn.	1	L10
L9	US-6327590-B1.did.	1	L9
	DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR		
L8	meta near search near engines	82	L8
	DB=USPT; PLUR=YES; OP=OR		
L7	5848397.pn.	1	L7
L6	5903882.pn.	1	L6
L5	5918014.pn.	1	L5
L4	5920859.pn.	1	L4
	DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR		

<u>L3</u>	5864845.uref.	15	<u>L3</u>
<u>L2</u>	5826261.pn.	2	<u>L2</u>
<u>L1</u>	5864845.pn.	2	<u>L1</u>

END OF SEARCH HISTORY

بحث على الإنترنت

Search Tools	

Search Tools

Search Engines & Spiders

Alta Vista

Excite

HotBot

Infoseek

LookSmart

Lycos

Northern Light

Popular Directories

Yahoo

Internet Public Library

Popular Meta Search Engines

MetaCrawler

Internet Sleuth

Dogpile

Mamma

Northernlight

Answers Searching

Ask Jeeves

Information Please

Specialized Search Tools

Liszt

Deja News

News Services

NewsBot

NewsHub

News Index

Search Tools for Kids

Ask Jeeves For Kids

Disney Internet Guide

Lycos SafetyNet

Yahooligans

Alta Vista	Consistently the largest search engine on the web, in terms of pages indexed, and is a particular favorite among researchers. Try their translation feature. Refine function allows narrowing of search by requiring or excluding topics from initial search.0.. Take the time to learn how to use advanced searches.
Excite	Excite Search taps into the traditional search engine listings, created by crawling the web. Channels By Excite lists sites by topics. These sites have been approved by editors, and sometimes also have reviews. There is also much associated subject information, discussion areas and more. "More Like This" allows searcher to ask for more sites like individual ones retrieved.
HotBot	HotBot grew out of a parrallel-processing project at UC: Berkeley. Its pull-down menus make complex searching easier. Demonstrates first-rate speed and allows searches for domain names and searches by media type.
Infoseek	Allows narrowing of

	search and searches by URL, ~~es,~~ links etc. "Channels" inlcude Education and Careers. "Worth a click" feature links to interesting sites and current events.
LookSmart	Type in your query and Search will first find sites selected and reviewed by our editors, next it will hunt in AltaVista's index. Explore speeds you through familiar categories to help quickly pinpoint web destinations tailored to your interests.
Lycos	Comprehensive. You can choose to search only Lycos Top 5%, a directory of sites by Lycos reviewers. Offers links and selected Web sites under many categories.
Northern Light	Simple, Straightforward. Search results are organized into folders in which the returns are classified by subject, type, source and language.
Popular Directories Directories are compiled by staff who select and organize Web sites by topic. If you like to browse for information, start with one of these.	
Yahoo	Clear organization and broad coverage. A great place to get comfortable with the web.
Internet Public Library	Serves the public by finding, evaluating, selecting, organizing, describing, and creating quality information resources. A great resource.

Popular Meta Search Engines

Meta-searchers send your query to many search engines at the same time.

Unlike search engines, metacrawlers don't crawl the web themselves to build listings. Instead, they allow searches to be sent to several search engines all at once. The results are then blended together onto one page. Below are some of the major metacrawlers.

MetaCrawler	One of the oldest meta search services, MetaCrawler began in July 1995 at the University of Washington. MetaCrawler was purchased by go2net, an online content provider, in Feb. 97. The commercial backing has helped improve the responsiveness of the service.
Internet Sleuth	Allows you to search the standard search engine choices or a huge number of specialty sites, all from the same place.
Dogpile	Sends a search to a customizable list of search engines, directories and specialty search sites.
Mamma	Mamma offers simultaneous coverage of the major search engines in one simple query. This is why the program is called "Mamma". Mamma will also allow users to use any type of search syntax so you don't need to learn a special syntax. Moreover, Mamma can include special syntax by itself for queries that are not correct. Sends search requests to six major search engines. Mamma

uses: Alta Vista, Excite, Infoseek, Lycos, WebCrawler and Yahoo.	
Northernlight	Part of it is a Web search engine; part of it is a full-text database.
Answers Searching	
Ask Jeeves	Ask Jeeves is a human-powered search service that aims to direct you to the exact page that answers your question. If it fails to find a match within its own database, then it will provide matching web pages from various search engines.
Information Please	Information Please almanacs are favorites among researchers who need trustworthy facts. This site allows searching across Information Please's various almanacs, its encyclopedia and its dictionary.
Specialized Search Tools	
Liszt	Long a favorite for those looking for mailing lists.
Deja News	Deja News is devoted to searching newsgroup discussions, with archives stretching back to March 1995. Anyone who's ever struggled to find a relevant newsgroup for a particular topic by looking at newsgroup names will find Deja News an incredible resource.
News Services These services provide exceptionally good results for current event searching, because they crawl only news sites once or twice a day. Thus, the results are usually focused and timely.	

NewsBot	HotB⬤ news-only search service.
NewsHub	Search results are powered by News Index (below). News can also be browsed by topic.
News Index	Launched in April 1996, indexes news stories from hundreds of sources, worldwide. The goal is to refresh the index once per hour.
Search Tools for Kids	
Ask Jeeves For Kids	Ask Jeeves is a unique service where you enter a question, and Ask Jeeves tries to point you to the right web page that provides an answer. At Ask Jeeves For Kids, answers have been vetted for appropriateness. Also, if Ask Jeeves cannot answer a question, it pulls results from various search engines in its metacrawler mode. At Ask Jeeves For Kids, no site that is on SurfWatch's block list will be listed.
Disney Internet Guide (DIG)	Disney's kids' guide to the Internet, which contains only sites considered appropriate for children.
Lycos SafetyNet	Allows parents to screen possibly objectionable sites from Lycos search results. Unlike most of the other services listed, this means that searches can be done across the entire web, as opposed to among a set of chosen sites. That is helpful for those doing research on obscure topics, or those who simply like Lycos

	result t who are concerned that undesirable sites may appear.
Yahooligans	Yahoo for kids, designed for ages 7 to 12. Sites are hand-picked to be appropriate for children. Also, unlike normal Yahoo, searches will not be forwarded to Yahoo's search engine partner Inktomi if there is no match from within the Yahooligan listings. This prevents possibly objectionable sites from slipping onto the screen. Additionally, adult-oriented banner advertising will not appear within the service. Yahooligans is the oldest major directory for children, launched in March 1996.

مواقع مصرية   بحث على الإنترنت   إتصل بنا

Refine Search

Search Results -

Terms	Documents
L17 and (estimate or calculate or rank) and coverage	15

Database:

US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Search:

[Refine Search]

[Recall Text]   [Clear]   [Interrupt]

Search History

DATE: Friday, January 30, 2004   Printable Copy   Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L19	L17 and (estimate or calculate or rank) and coverage	15	L19
L18	L17 and meta near inform$	7	L18
L17	(meta adj search adj engines or multiple near party near search near engines or multi-party near search near engines)	78	L17
L16	709/245	1839	L16
L15	709/227	3482	L15
L14	709/224	4818	L14
L13	709/218	3038	L13
L12	709.clas.	26282	L12
L11	345/968	277	L11
L10	345/866	428	L10
L9	345.clas.	64799	L9
L8	707/10	7626	L8

L7	707/7	1381	L7
L6	707/6	2281	L6
L5	707/4	3332	L5
L4	707/2	3559	L4
L3	707/5	2723	L3
L2	707/3	5714	L2
L1	707.clas.	18699	L1

END OF SEARCH HISTORY

Freeform Search

Database:
US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term:

Display: `10` **Documents in Display Format:** `-` **Starting with Number** `1`

Generate: ○ Hit List ◉ Hit Count ○ Side by Side ○ Image

[Search] [Clear] [Interrupt]

Search History

DATE: Friday, January 30, 2004   Printable Copy   Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=USPT; PLUR=YES; OP=OR			
L27	0408296.pn.	1	L27
L26	5566330.pn.	1	L26
L25	5708825.pn.	1	L25
L24	5848410.pn.	1	L24
L23	5941944.pn.	1	L23
L22	5941944.pn.	1	L22
L21	6282533.pn.	1	L21
L20	6483702.pn.	1	L20
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L19	L17 and (estimate or calculate or rank) and coverage	15	L19
L18	L17 and meta near inform$	7	L18
L17	(meta adj search adj engines or multiple near party near search near engines or multi-party near search near engines)	78	L17
L16	709/245	1839	L16
L15	709/227	3482	L15

L14	709/224	4818	L14
L13	709/218	3038	L13
L12	709.clas.	26282	L12
L11	345/968	277	L11
L10	345/866	428	L10
L9	345.clas.	64799	L9
L8	707/10	7626	L8
L7	707/7	1381	L7
L6	707/6	2281	L6
L5	707/4	3332	L5
L4	707/2	3559	L4
L3	707/5	2723	L3
L2	707/3	5714	L2
L1	707.clas.	18699	L1

END OF SEARCH HISTORY

Google™

meta search engines<1997

Google Search

| **Web** | Images | Groups | Directory | News |

Searched the web for **meta search engines<1997**.   Results **31 - 40** of about **171,000**. Search took **0.38** seconds.

Open Directory - Computers: Internet: S arching: S arch Engines

... Also offering a **meta** tag builder and analyzer. ... (March, 2000);
Estimating the Relative
Size and Overlap of Public Web **Search Engines** - Research paper
by ... (**1997**). ...
dmoz.org/Computers/Internet/Searching/Search_Engines/ - 15k -
Cached - Similar pages

HTML META Tags

... Metadata Registries (July **1997**): Metadata **Search Engine**;
MetaWeb - the Australian
metadata project at DSTC; The Metadata Repository Service; **Meta**
Content Framework ...
Description: Taxonomy of HTML **meta** tags, with references.
Discusses using tags to change character sets, refresh...
Category:
Computers > Data Formats > ... > HTML > Tutorials > Meta Tags
vancouver-webpages.com/META/metatags.detail.html - 30k - Cached
- Similar pages

[PDF] Adaptively Constructing the Query Interface for **Meta-Search** ...

File Format: PDF/Adobe Acrobat - View as HTML
... A **meta-search engine** based on an adaptive constraints-based query interface model
will ... Interests, in Proceedings of CHI '97 (Atlanta, GA, April **1997**), ACM Press ...
www.iuiconf.org/01pdf/2001-002-0016.pdf - Similar pages

SEARCH ENGINE/ META TAG INFO

... Watch How **Search Engines** Rank Web Pages How To Use **Meta** Tags **Meta** tags- What, Where,
When, Why? Return to the INTERNET/ HTML/ SHAREWARE. Since 7-01-**1997** Updated ...
www.chiro.org/LINKS/metatag.shtml - 5k - Cached - Similar pages

Chemie.DE **Search Engine**

... Home | **Search Engine** | **Meta Search** | Conferences | Departments ... **1997**-2004 Chemie.DE
Information Service GmbH a Life Science Network Division www.Chemie.DE ...
www.chemie.de/search/?language=e - 29k - Cached - Similar pages

HotSource HTML Help - HTML - **Meta** Tags

... is mostly used to help **search engines** locate information ... so it can be displayed for
people **searching**. ... **meta** http-equiv="Copyright" content="holder name -- **1997** ...
www.sbrady.com/hotsource/html/meta.html - 8k - Cached - Similar pages

META - **Meta**-information

... If you insert a keyword more than seven times here, the whole tag will be ignored!
<**META** NAME="description" CONTENT="This is a site"> **Search engines** which ...
www.htmlhelp.com/reference/wilbur/head/meta.html - 6k - Cached - Similar pages

Let them **search engine** robots know what your page is really about ...

... to find everything about blue marbles!"> <**META** NAME="keywords ... Created on 22-Jan-**1997**. ...
www.chami.com/tips/internet/012297l.html - 19k - Cached - Similar pages

Free Pint Portal - Industry Research

... Metasearch - **Meta search engine** [11/05/00 ... URL **Search Engine** - **Search** for words appearing

www.freepint.com/portal/industry/ indu⬤.php3?category_id=185 - 48k - Jan 28, 2⬤- <u>Cached</u> - <u>Similar pages</u>

<u>SocioSite</u>: **META SEARCH ENGINES**
Debriefing A **meta s arch engine** written entirely with Java ... Excellent multi-threaded
search engine that combines ... is that poorly constructed **searches** can take ...
www2.fmg.uva.nl/sociosite/search/Search3.html - 12k - <u>Cached</u> - <u>Similar pages</u>

◀ Goooooooooooogle ▶
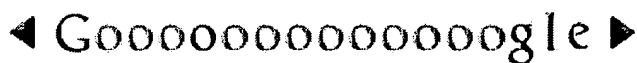
Result Page: **Previous** <u>1</u> <u>2</u> <u>3</u> 4 <u>5</u> <u>6</u> <u>7</u> <u>8</u> <u>9</u> <u>10</u><u>11</u><u>12</u><u>13</u>   **Next**

| meta search engines<1997 | Google Search | Search within results |

<u>Google Home</u> - <u>Advertise with Us</u> - <u>Business Solutions</u> - <u>Services & Tools</u> - <u>Jobs, Press, & Help</u>

©2004 Google

HOME

Chami.com Tips

Home > Chami.com Tips > Internet > Let them search engine robots know what...

Advertisement

Click for Details

Let them search engine robots know what your page is really about using DESCRIPTION and KEYWORDS meta tags

Search

Sponsored ↓

Pickaweb UK

Domain Names

Web Design UK

Web Hosting UK

Cheap Domain Name Registration

Merkaweb

Dominios

Business Web Hosting

Click for Details

Keywords

HTML Mini Tutorial

World Wide Web

Whenever you submit your WWW address to an automated search engine robot, it will determine which keywords and descriptions to use for your pages. Of course, robots aren't "human enough" to determine the best way to describe your pages nor find the best keywords to assign to your pages. Here's how to provide your own descriptions and keywords for search engines to use:

All you have to do is place two META tags named "description" and "keywords" in-between your <HEAD> and </HEAD> tags:

```
<HTML>

<HEAD>

<META NAME="description"
   CONTENT="Come here to find
everything about blue marbles!">

<META NAME="keywords"
   CONTENT="blue marbles,
   marbles, marble information">

</HEAD>

<BODY>

<...rest of your page...>

</BODY>
</HTML>
```

The text colored in green of course should be changed to represent your page; don't forget to separate your keywords using commas.

See Also

- Don't cache my page!
- Automatically redirect your visitors to your new home page.
- Keeping robots, spiders and wanderers away from your site using robots.txt, meta tags and other methods

Newsletter »

Applicable Keywords: *HTML, Mini Tutorial, World Wide Web*

enter email
Subscribe

The Internet Public Library

www.ipl.org "Especially for Librarians" Section

STUMPERS-L

www.cuis.edu/~stumpers/intro.html

The Stumpers-L electronic mailing list official home page. Sponsored by the
Graduate School of Library and Information Science at Dominican University, Stumpers-L
was founded as an email-based resource where reference librarians can help each other find
the answers to difficult questions.

Always remember to search the archives before submitting a request!

Your question may already have an answer.

Special Tools for Librarians - Continued

PUBLIB-NET is an electronic discussion list, or listserv, concerned with the use of the Internet in public libraries. (It is actually a subset of the listserv, **PUBLIB**, which discusses all issues related to public libraries.) Issues discussed include connectivity, public access to the Internet, user and staff training, resources of interest to public librarians (online, print, video, other), electronic freedoms and responsibilities, new technologies for public library Internet access, National and regional public telecommunications policy and public libraries, and more. Messages are sent to subscribers once per day in "digest" form. Both PUBLIB-NET and PUBLIB have searchable archives of previous postings.

To subscribe to the list and receive messages posted to it:

Send an e-mail message to:

listserv@sunsite.berkeley.edu

Leave the subject line blank. In the body of the message, type:

subscribe PUBLIB-NET yourfirstname yourlastname

(using, of course, your own first and last name)

To post messages to the list:

Use the e-mail address:

PUBLIB-NET@ sunsite.berkeley.edu

To unsubscribe from the list:

Send a message to:

LISTSERV@sunsite.berkeley.edu

Leave the subject line blank. In the body of the message, type:

signoff PUBLIB-NET

To search the archives of previous postings to PUBLIB and PUBLIB-NET:

connect to: *sunsite.berkeley.edu/PubLib/archive.html*

www.dejanews.com

Newsgroups - One of Usenet's huge collection of topic groups or fora. Usenet groups can be 'unmoderated' (anyone can post) or 'moderated'(submissions are automatically directed to a moderator, who edits or filters and then posts the results). Some newsgroups have parallel mailing lists for Internet people with no netnews access, with postings to the group automatically propagated to the list and vice versa. Some moderated groups (especially those which are actually gatewayed Internet mailing lists) are distributed as 'digests', with groups of postings periodically collected into a single large posting with an index.

Note that words are commonly misspelled in usenet archives. If you have what you think may be a misspelled word, search DejaNews and you may find the correct spelling in a thread!

See the DejaNews Help Wizard at www.dejanews.com/help/wizard.shtml for more information.

Liszt, the mailing list directory
www.liszt.com

Liszt helps you find mailing lists that might interest you. Then it tells you how to get more information, and how to join. But you'll still have to use your e-mail program to actually join and read the group. So Liszt just gets you started, by giving you instructions on how to do this.

What do I Need to Know about Java, Helper Apps, and Plug-ins?

JavaScript and Java are advanced technologies that software developers and page authors use to enhance the delivery of Internet information. From the viewpoint of typical users, these technologies are transparent, built into the system of Internet servers, applications, and content. You can take advantage of the technologies with no effort on your part.

A **helper application** is a separate stand-alone software program with capabilities which Netscape does not possess. Examples are Stuffit Expander for the Macintosh and PKUnzip for Windows, both of which decompress downloaded files. You can also download a number of helper applications from a Web page maintained by Netscape Communications. **home.netscape.com/assist/helper_apps**

A **plug-in** is software that works inside Netscape to extend its capabilities. You can only access a plug-in from within Netscape. You do not have to do anything to configure Netscape to use a plug-in aside from installing it correctly; Netscape uses plug-ins just like a built-in capability. You can also download a number of plug-ins from a Web page maintained by Netscape Communications. **www.netscape.com/comprod/products/navigator/version_2.0/plugins/index.html**

Acrobat Reader Tutorial - A step-by-step tutorial on how to use
Acrobat Reader to view PDF (Portable Document Format) documents.
w3.aces.uiuc.edu/AIM/scale/tutorials/Acrobat/index.html

Acrobat
Reader
tutorial

Downloading Adobe Acrobat Reader Software:
www.adobe.com/prodindex/acrobat/readstep.html

Envoy Viewer - Information and Download
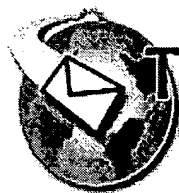
TUMBLEWEED SOFTWARE

Leading Solutions for Internet Delivery

www.twcorp.com/viewer.htm

How to Download Adobe Acrobat Reader

Adobe Acrobat Reader

The free Adobe(R) Acrobat(R) Reader allows you to view, navigate, and print PDF files across all major computing platforms. Acrobat Reader is the free viewing companion to Adobe Acrobat 3.0 and to Acrobat Capture(R) software.

Download the free Adobe Acrobat Reader by following these easy steps:

1. **Point your browser to the Adobe Reader download page**

   www.adobe.com/prodindex/acrobat/readstep.html

2. **Register with Adobe (if you haven't already registered as an Acrobat Reader user).**

3. **Choose the Reader version, platform version, and language version you need from the pop-up lists listed on the website.**

For additional information, click on the links provided by Adobe.

Evaluating Web Sources

Adapting Five Traditional Print Evaluation Criteria to Web Resources

#1: Accuracy of Web Resources

        Almost anyone can publish on the Web
        Many Web resources not verified by editors and/or fact checkers
        Web Standards to ensure accuracy yet to be fully developed

#2: Authority of Web Resources

        Often difficult to determine authorship of Web Sources
        If author's name is listed, his/her qualifications frequently absent
        Publisher responsibility often not indicated

#3: Objectivity of Web Resources

        Goals/aims of persons or groups presenting material often not clearly stated
        Web often functions as a "virtual soapbox"

#4: Currency of Web Resources

        Dates not always included on Web pages
        If included, a date may have various meanings:

                Date information first written
                Date information placed on Web
                Date information last revised

#5: Coverage of Web Resources

        Web coverage may differ from print coverage
        Often hard to determine extent of Web coverage

Evaluating Web Sources - Continued

Applying Evaluation Techniques to Specific Types of Web Resources

Step 1: Identify the Type of Web Page
1. Entertainment
2. Business/Marketing
3. Reference/Informational
4. News
5. Advocacy
6. Personal Page

Step 2: Use the Appropriate Checklist

Step 3: Based on the Checklist Criteria, Determine the Relative Quality of the Web Page
The more "yes" answers to questions indicates a higher quality Web page

Conclusion: Remember!

The Web is only one source of information
1. It can be very useful for researching certain topics
2. It can be almost useless for other topics
3. To research a topic thoroughly, use a variety of sources both Web and non-Web
Web evaluation techniques are just beginning to be developed
Technology is outpacing ability to create standards and guidelines
Establishing evaluation procedures will be an ongoing evolutionary process

URL: www.science.widener.edu/~withers/evalout.htm

Index

$960.00   250 $3.84 each

Search Engines, Directories and Metacrawlers

For your convenience we list here some of the most usefull Search Engines, Directories and Metacrawlers available. Most data were taken from the excelent **Search Engine Watch** site (*in particular their Search Engine Listings section, where you will find even more links*).

- Major Search Engines and Directories (*ordered alphabeticaly*)

AltaVista
http://www.altavista.com/

   AltaVista is consistently one of the largest search engines on the web, in terms of pages indexed. Its comprehensive coverage and wide range of power searching commands makes it a particular favorite among researchers. It also offers a number of features designed to appeal to basic users, such as "Ask AltaVista" results, which come from Ask Jeeves (see below), and directory listings primarily from the Open Directory. AltaVista opened in December 1995. It was owned by Digital, then run by Compaq (which purchased Digital in 1998), then spun off into a separate company which is now controlled by CMGI.

AOL Search
http://search.aol.com/

   AOL Search allows its members to search across the web and AOL's own content from one place. The "external" version, listed above, does not list AOL content. The main listings for categories and web sites come from the Open Directory (see below). Inktomi (see below) also provides crawler-based results, as backup to the directory information. Before the launch of AOL Search in October 1999, the AOL search service was Excite-powered AOL NetFind.

Ask Jeeves
http://www.askjeeves.com/

   Ask Jeeves is a human-powered search service that aims to direct you to the exact page that answers your question. If it fails to find a match within its own database, then it will provide matching web pages from various search engines. The service went into beta in mid-April 1997 and opened fully on June 1, 1997. Results from Ask Jeeves also appear within AltaVista.

Direct Hit
http://www.directhit.com/

   Direct Hit is a company that works with other search engines to refine their results. It does this by monitoring what users click on from the results they see. Sites that get clicked on more than others rise higher in Direct Hit's rankings. Thus, the service dubs itself a "popularity engine." Direct Hit's technology is currently best seen at HotBot. It also refines results at Lycos and is available as an option at LookSmart and MSN Search. The company also crawls the web and refines this database, which can be viewed via the link above.

Excite
http://www.excite.com/

   Excite is one of the most popular search services on the web. It offers a medium-sized index and integrates non-web material such as company

information and sports scores into its results, when appropriate. Excite was launched in late 1995. It gre~~w qu~~ickly in prominence and consumed ~~two~~ of its competitors, Magellan in July 1996, and WebCrawler in November 1996. These continue to run as separate services.

FAST Search
http://www.alltheweb.com/

Formerly called All The Web, FAST Search aims to index the entire web. It was the first search engine to break the 200 million web page index milestone. The Norwegian company behind FAST Search also powers the Lycos MP3 search engine. FAST Search launched in May 1999.

Go / Infoseek
http://www.go.com/

Go is a portal site produced by Infoseek and Disney. It offers portal features such as personalization and free e-mail, plus the search capabilities of the former Infoseek search service, which has now been folded into Go. Searchers will find that Go consistently provides quality results in response to many general and broad searches, thanks to its ESP search algorithm. It also has an impressive human-compiled directory of web sites. Go officially launched in January 1999. It is not related to GoTo, below. The former Infoseek service launched in early 1995.

Google
http://www.google.com/

Google is a search engine that makes heavy use of link popularity as a primary way to rank web sites. This can be especially helpful in finding good sites in response to general searches such as "cars" and "travel," because users across the web have in essence voted for good sites by linking to them.

GoTo
http://www.goto.com/

Unlike the other major search engines, GoTo sells its main listings. Companies can pay money to be placed higher in the search results, which GoTo feels improves relevancy. Non-paid results come from Inktomi. GoTo launched in 1997 and incorporated the former University of Colorado-based World Wide Web Worm. In February 1998, it shifted to its current pay-for-placement model and soon after replaced the WWW Worm with Inktomi for its non-paid listings. GoTo is not related to Go (Infoseek).

HotBot
http://www.hotbot.com/

Like AltaVista, HotBot is another favorite among researchers due to its large index of the web and many power searching features. In most cases, HotBot's first page of results comes from the Direct Hit service (see above), and then secondary results come from the Inktomi search engine, which is also used by other services. It gets its directory information from the Open Directory project (see below). HotBot launched in May 1996 as Wired Digital's entry into the search engine market. Lycos purchased Wired Digital in October 1998 and continues to run HotBot as a separate search service.

Inktomi
http://www.inktomi.com/

Originally, there was an Inktomi search engine at UC Berkeley. The creators then formed their own company with the same name and created a new Inktomi index, which was first used to power HotBot. Now the Inktomi index also powers several other services. All of them tap into the same index, though results may be slightly different. This is because Inktomi provides ways for its partners to use a common index yet distinguish themselves. There is no way to query the Inktomi index directly, as it is only made available through Inktomi's

IW n
http://www.iwon.com/

Backed by US television network CBS, iWon has a directory of web sites generated automatically by Inktomi, which also provides its more traditional crawler-based results. iWon gives away daily, weekly and monthly prizes in a marketing model unique among the major services. It launched in Fall 1999.

LookSmart
http://www.looksmart.com/

LookSmart is a human-compiled directory of web sites. In addition to being a stand-alone service, LookSmart provides directory results to MSN Search, Excite and many other partners. AltaVista provides LookSmart with search results when a search fails to find a match from among LookSmart's reviews. LookSmart launched independently in October 1996, was backed by Reader's Digest for about a year, and then company executives bought back control of the service.

Lycos
http://www.lycos.com/

Lycos started out as a search engine, depending on listings that came from spidering the web. In April 1999, it shifted to a directory model similar to Yahoo. Its main listings come from the Open Directory project, and then secondary results come from either Direct Hit or Lycos' own spidering of the web. In October 1998, Lycos acquired the competing HotBot search service, which continues to be run separately.

MSN Search
http://search.msn.com/

Microsoft's MSN Search service is a LookSmart-powered directory of web sites, with secondary results that come from AltaVista. RealNames and Direct Hit data is also made available. MSN Search also offers a unique way for Internet Explorer 5 users to save past searches.

Netscape Search
http://search.netscape.com/

Netscape Search's results come primarily from the Open Directory and Netscape's own "Smart Browsing" database, which does an excellent job of listing "official" web sites. Secondary results come from Google. At the Netscape Netcenter portal site, other search engines are also featured.

Northern Light
http://www.northernlight.com/

Northern Light is another favorite search engine among researchers. It features one of the largest indexes of the web, along with the ability to cluster documents by topic. Northern Light also has a set of "special collection" documents that are not readily accessible to search engine spiders. There are documents from thousands of sources, including newswires, magazines and databases. Searching these documents is free, but there is a charge of up to $4 to view them. There is no charge to view documents on the public web -- only for those within the special collection. Northern Light opened to general use in August 1997.

Open Direct ry
http://dmoz.org/

The Open Directory uses volunteer editors to catalog the web. Formerly known as NewHoo, it was launched in June 1998. It was acquired by Netscape in November 1998, and the company pledged that anyone would be able to use information from the directory through an open license arrangement. Netscape

itself was the first .lnoeusee...Lycos and AOL Search also make heavy use of Open Directory data, while AltaVista and HotBot prominently feature Open Directory categories within their results pages.

RealNames
http://www.realnames.com/

The RealNames system is meant to be an easier-to-use alternative to the current web site addressing system. Those with RealNames-enabled browsers can enter a word like "Nike" to reach the Nike web site. To date, RealNames has had its biggest success through search engine partnerships. In particular, it is strongly featured in results at AltaVista, Go and MSN Search.

Snap
http://www.snap.com/

Snap is a human-compiled directory of web sites, supplemented by search results from Inktomi. Like LookSmart, it aims to challenge Yahoo as the champion of categorizing the web. Snap launched in late 1997 and is backed by Cnet and NBC.

WebCrawler
http://www.webcrawler.com/

WebCrawler has the smallest index of any major search engine on the web -- think of it as Excite Lite. The small index means WebCrawler is not the place to go when seeking obscure or unusual material. However, some people may feel that by having indexed fewer pages, WebCrawler provides less overwhelming results in response to general searches. WebCrawler opened to the public on April 20, 1994. It was started as a research project at the University of Washington. America Online purchased it in March 1995 and was the online service's preferred search engine until Nov. 1996. That was when Excite, a WebCrawler competitor, acquired the service. Excite continues to run WebCrawler as an independent search engine.

Yahoo
http://www.yahoo.com/

Yahoo is the web's most popular search service and has a well-deserved reputation for helping people find information easily. The secret to Yahoo's success is human beings. It is the largest human-compiled guide to the web, employing about 150 editors in an effort to categorize the web. Yahoo has over 1 million sites listed. Yahoo also supplements its results with those from Inktomi. If a search fails to find a match within Yahoo's own listings, then matches from Inktomi are displayed. Inktomi matches also appear after all Yahoo matches have first been shown. Yahoo is the oldest major web site directory, having launched in late 1994.

- Major Metacrawlers

Go2Net / MetaCrawler
http://www.go2net.com/

One of the oldest meta search services, MetaCrawler began in July 1995 at the University of Washington. MetaCrawler was purchased by go2net, an online content provider, in Feb. 97. The commercial backing has helped improve the responsiveness of the service. MetaCrawler now powers searches at the Go2Net portal site.

SavvySearch
http://www.savvysearch.com/

Another one of the older metasearch services, around since May 1995 and formerly based at Colorado State University. It is highly customizable and

covers a huge range of general and specialty search sites.

D gpile
http://www.dogpile.com/

Popular metasearch site that sends a search to a customizable list of search engines, directories and specialty search sites. Dogpile also runs the MetaFind metasearch site that sends searches only to crawler-based search engines.

Inference Find
http://www.infind.com/

An alternative to typical metacrawlers, Inference lists results grouped by subject, rather than by search engine or in one giant list. For example, a search for "Uma Thurman" groups results into "Uma Thurman" and "Pulp Fiction," among other categories. It taps into Alta Vista, Excite, Infoseek, Lycos, WebCrawler and Yahoo. The service began in May 1995, moving to its present domain in Oct. 1996.

ProFusion
http://www.profusion.com/

Customizable, with broken link detection available. Formerly based at the University of Kansas.

Mamma
http://www.mamma.com/

Sends search requests to major search services.

The Big Hub
http://www.thebighub.com/

Allows you to search many major search engines or a huge number of specialty sites, all from the same place. Formerly the Internet Sleuth.

C4
http://www.c4.com/

C4 allows meta searching against several major search engines, with a nice, clean interface.

- Multimedia Search Engines (*General*)

AltaVista Photo Finder
http://image.altavista.com/cgi-bin/avncgi

Impressive service that lets you find photos, images, audio and video clips from all over the web. Search results feature thumbnails of images found.

Ditto
http://www.ditto.com/

Search or browse to find images on the web. Matches are displayed in thumbnail format. Formerly known as Arriba Vista.

Lyc s Pictures and S unds
http://www.lycos.com/picturethis/

The Lycos multimedia search service. It features images organized by category, from the PicturesNow catalog. You can browse categories and view thumbnails of these pictures. Search mode lets you scan the web for pictures or sounds of interest, but no thumbnails are provided.

Sc ur.Net
http://scour.net/

. A multimedia search engine that allows users to find audio, video and images on the web, including MP3 files.

StreamSearch.c m
http://www.streamsearch.com/

  Directory of multimedia resources on the web. Search or browse categories.

MIDI Expl rer
http://www.musicrobot.com/

  Allows you to search for MIDI files.

- Multimedia Search Engines (*MP3*)

Lycos / FAST MP3 Search
http://mp3.lycos.com/

  Over 1/2 million MP3 files are listed here, in an index that's updated on an hourly basis. The freshest, most dependable links are listed first.

MP3.com
http://mp3.com/

  All things about MP3, including thousands of legal MP3 files.

MP3meta
http://www.mp3meta.com/

  Search all the major MP3 search engines at once through this metasearch service from SavvySearch.

2Look4
http://www.2look4.com/

  Has an option to filter out unreliable sites from your MP3 searches.

AudioGalaxy
http://www.audiogalaxy.com/

  Displays site speed and reliability information for each match.

Oth.net
http://oth.net/

  A bare-bones interface, but comprehensive coverage of many files.

Audiofind
http://www.audiofind.com/

  Browse by artist or genre, or keyword search for MP3 files.

MediaLeech search
http://medialeech.m4d.com/

  MP3 search engine.

Arianna MP3
http://mp3.iol.it

  Italian MP3 search engine allowing you to search by artist, song title or album title.

Gets ngs - MP3 Search Engines
http://altern.org/getsongs

  Query multiple MP3 search engines from one place.

Manic Music
http://www.m-music.net

  Lets you choose to search from many MP3 search engines, though only one at a time.

S undcrawler
http://www.soundcrawler.com/

- News Search Engines

Excite NewsTracker
http://nt.excite.com/

  Excite's news-only search service. This is a personal favorite, especially for the way you can train the service to learn what you like by news topic.

News Index
http://www.newsindex.com/

  Indexes news stories from hundreds of sources, worldwide. The goal is to refresh the index once per hour. Launched in April 1996.

HotBot News Search / NewsBot
http://www.newsbot.com/

  HotBot's news-only search service.

Northern Light's Current News
http://www.northernlight.com/news.html

  Information in the index is gathered from various news sources, such as the Associated Press and Business Wire. Content is constantly refreshed throughout the day. Options include the ability to sort news by date or relevance, and to narrow searches to within predefined categories and timespans.

NewsHub
http://www.newshub.com/

  News from a variety of sources worldwide.

NewsTrawler
http://www.newstrawler.com/

  Allows you to send a query to one or more news sites from one location. Hundreds of sites are listed, by country and by category.

Paperball
http://www.paperball.de/

  Produced by the German Fireball search service, Paperball lets you search for German news.

Paperboy
http://www.paperboy.de/

  Covers newspapers and other selected media from Germany, as well as worldwide.

NewsN w
http://www.newsnow.co.uk/

  Search up to 30 days worth of headlines from nearly 150 news sources. Especially aimed at UK users.

- **1stHeadlines**
http://www.1stheadlines.com/

  News search engine that includes local, regional, national and international news sources.

AltaVista Canada - Canadian News Index
http://www.altavistacanada.com/

  The Canadian News Index gathers information from over 300 Canadian news sources daily. To use the service, simply select the appropriate option on the AltaVista Canada homepage. It appears just above the search box.

Fanagalo
http://www.fanagalo.co.za/

  A news search engine, Fanagalo crawls web sites that have news from a South African perspective. Information is updated daily.

InfoJump
http://www.infojump.com/

  Search indexed articles from over 4,000 electronic publications.

TotalNews
http://www.totalnews.com/

- Specialty Search Engines

Too many to list here. See the link above.

- Regional Search Engines

For a long listing of regional search engines see the link above.
Some Greek (*or Greek-related*) search engines are:

The Greek Explorer And Indexer
http://www.hiway.gr/

  The Greek Explorer appears to be a Greek domain crawler, while the Indexer is a directory of sites.

Webindex Internet Search - Greece
http://www.webindex.gr/

  Search engine that crawls Greek domains, plus an associated directory. It launched in Jan. 1997.

iBoom
http://www.iboom.com/

  Directory of Greek and Greek-related web sites.

IN.GR
http://www.in.gr/

  Directory of Greek web sites.

R BBy
http://www.robby.gr/

Search engine that crawls Greek domains, plus an associated directory. It
launched in Nov. 1997. It also powers the X-treme content service.

G Greece.com
http://www.gogreece.com/

A human indexed directory service.

Greek Internet Direct ry
http://www.directory.gr/

A human indexed directory service.

HR-Net Interesting Nodes Collection
http://www.hri.org/nodes/greece.html

A human indexed directory service.

HACK.gr Meta Search Engine
http://www.hack.gr/mse/

A metacrawler for the Greek Cyberspace.

Phantis
http://www.phantis.gr/

A search engine. It launched in May 1997.

Thea
http://www.thea.gr/

Search engine that crawls Greek domains, plus an associated directory.

EuroSeek
http://www.euroseek.net/page?ifl=gr

The Greek version of the EuroSeek search engine.

HELLAS MAP
http://www.forthnet.gr/hellas/

An indexed directory service. Entries are also indexed by geographic location.
It launched in Sep. 1995.

[ΑΝΑΖΗΤΗΣΗ] [ΕΥΡΕΤΗΡΙΟ ΣΕΛΙΔΩΝ] [ΕΠΙΚΟΙΝΩΝΙΑ] [ΒΟΗΘΕΙΑ & ΠΛΗΡΟΦΟΡΙΕΣ] [ΑΡΧΙΚΗ ΣΕΛ